MESTRADO EM

MÉTODOS QUANTITATIVOS PARA A DECISÃO ECONÓMICA E EMPRESARIAL

TRABALHO FINAL DE MESTRADO

PROJETO

UTILIZAÇÃO DE ALGORITMOS DE APRENDIZAGEM AUTOMÁTICA PARA PREVISÃO DO SUCESSO DE FILMES

JOÃO ANTÓNIO VIEIRA PINHEIRO

ORIENTAÇÃO:

PROF.º DOUTOR CARLOS J. COSTA

OUTUBRO-2023

Agradecimentos

Em primeiro lugar gostaria de agradecer ao ISEG, que nos últimos anos, foi a minha segunda casa. Foi nesta casa que tive o privilégio de conhecer pessoas excelentes que espero poder levar para a vida.

Devo também um agradecimento ao Professor Carlos Costa por toda a sua disponibilidade, apoio, críticas e sugestões dadas que foram fundamentais para a elaboração de este Trabalho Final de Mestrado.

Agradeço também à minha família por estar sempre disponível para me ajudar. Em especial aos meus pais, por toda a força e apoio que me deram e principalmente por acreditarem em mim.

Resumo

Um dos objetivos mais importantes da análise de dados foca-se cada vez mais na aplicação de análises estatísticas para tentar prever a reação da sociedade a um novo produto. A indústria cinematográfica é uma das maiores e mais importantes indústrias de entretenimento do mundo, gerando milhares de milhões de euros todos os anos. O objetivo deste projeto é criar um sistema que apoie a decisão na criação de um novo filme, usando técnicas de aprendizagem automática (*machine learning)* de forma a minimizar o risco do investimento. O sistema prevê o sucesso do filme baseado nas críticas recebidas pelo mesmo, tendo sido usados *datasets* fornecidos pelo IMDb para obter características antes e após lançamento dos filmes, para o treino e o teste dos algoritmos. Neste projeto o algoritmo de machine learning que obteve a melhor previsão foi o *Gradient Boosting Regressor* tendo apresentado uma precisão de aproximadamente 88%.

Palavras-chave: filme; aprendizagem automática; previsão; sucesso, python

Abstract:

One of the most important objectives of data analysis is applying statistical analysis to try to predict society's reaction to a new product. The film industry is one of the largest and most important entertainment industries in the world, generating billions of euros every year. The aim of this project is to create a system that supports the decision to create a new film, using machine learning techniques in order to minimize the risk of investment. The system predicts the success of the film based on the reviews it has received, using *datasets* provided by IMDb to obtain the pre- and post-release characteristics of the films, to train and test the algorithms. In this project, the algorithm with the best prediction power was Gradient Boosting Regressor, with an accuracy of around 88%.

Keywords: movie; machine learning; prevision; success; python

Índice

Agradecimentos I
Resumo II
Abstract: III
Índice de Tabelas V
Índice de Figuras VI
Capítulo 1: Introdução 1
1.1- Enquadramento 1
1.2- Objetivos 2
1.3- Estrutura 2
Capítulo 2: Revisão de literatura 4
Capítulo 3: Metodologia 7
Capítulo 4: Resultados 10
4.1- Identificação das necessidades da indústria *(Business Understanding)* 10
4.2- Compreensão dos dados *(Data Understanding)* 11
4.3- Preparação dos dados *(Data Preparation)* 12
4.4- Modelação *(Modeling)* 23
4.5- Avaliação *(Evaluation)* 26
4.6- Implementação *(Deployment)* 28
Capítulo 5: Discussão 30
Capítulo 6: Conclusões, Limitações e Trabalhos Futuros 31
Referências Bibliográficas 32
Anexo 36

Índice de Tabelas

Tabela 1- Modelos de previsão do sucesso de filmes: Objetivos, Algoritmos e Precisões........ 5
Tabela 2- title.basics.tsv.gz (https://datasets.imdbws.com/title.basics.tsv.gz)......................... 11
Tabela 3- name.basics.tsv.gz (https://datasets.imdbws.com/name.basics.tsv.gz).................. 11
Tabela 4- title.ratings.tsv.gz (https://datasets.imdbws.com/title.ratings.tsv.gz) 12
Tabela 5- title.crew.tsv.gz (https://datasets.imdbws.com/title.crew.tsv.gz)............................ 12
Tabela 6- Frequência de classificação por grupo 15
Tabela 7- Média das avaliações por intervalo temporal... 16
Tabela 8- Visão geral das variáveis dos filmes 20
Tabela 9- VIF das variáveis do df "Movie" 21
Tabela 10- Componentes Principais.. 22
Tabela 11- Algoritmos de machine learning e a sua precisão.. 24
Tabela 12- Análise swot da solução da previsão utilizando algoritmos de machine learning . 28

Índice de Figuras

Figura 1- Abordagem CRISP-DM (Chapman et al., 1999) 8
Figura 2- Formatos contidos no ”titleType” antes da limpeza 13
Figura 3- Formatos contidos no “titleType” após limpeza 13
Figura 4- Número de lançamento de novos filmes por ano 14
Figura 5- Distribuição de avaliações 15
Figura 6- Média das avaliações por intervalo temporal 16
Figura 7- Relação entre avaliação e duração em filmes de comédia 16
Figura 8- Relação entre avaliação e duração em filmes de ação 17
Figura 9- Wordcloud das palavras mais usadas nos títulos de filmes 18
Figura 10- Gráfico de barras com frequência de cada palavra 19
Figura 11- Matriz de correlação das variáveis 21
Figura 12- Regressão OLS 23
Figura 13- 5 fold cross-validation 26
Figura 14- Comparação entre os valores reais e previstos do modelo Gradient Boosting 27
Figura 15- Histograma de Resíduos 27

Capítulo 1: Introdução

1.1- Enquadramento

A indústria cinematográfica é uma das maiores e mais influentes formas de expressão artística em todo o mundo. A produção de filmes é uma atividade que envolve enormes investimentos de tempo e de recursos financeiros, e todos os anos milhares de filmes são produzidos, variando em qualidade. Surge então a questão fundamental: Como é que conseguimos saber se um filme é bom ou mau antes de o ver? Como é que decidimos que filme escolher para desfrutar e relaxar? Muitas vezes a nossa escolha é resultante das críticas disponíveis que encontramos na internet em sites como o IMDb ou o Metacritic. As avaliações nestes sites são uma grande referência para o público, representando a qualidade do conteúdo. Neste contexto, a capacidade de prever o potencial sucesso de um filme antes do seu lançamento pode ser uma ferramenta inestimável para estúdios, produtores e investidores. No entanto, o processo de previsão do sucesso de um filme é complexo, pois envolve um vasto leque de fatores, desde o elenco e o género até à qualidade do argumento e à estratégia de marketing. É neste contexto que surge a análise de dados e a utilização de algoritmos de aprendizagem automática, permitindo a exploração de tendências ocultas e a identificação de variáveis-chave que têm impacto na recetividade do público.

O IMDb (Internet Movie Database), é uma das maiores e mais abrangentes fontes de informação sobre filmes, oferecendo uma rica fonte de dados que pode ser explorada. Os *datasets* do IMDb contêm uma vasta quantidade de informação sobre cada filme, incluindo detalhes sobre o elenco, género, classificação, número de críticas, entre outros. Utilizando esses dados é possível criar modelos de machine learning que podem identificar os principais fatores que contribuem para o sucesso de um filme. É fundamental destacar que o IMDb fornece informações detalhadas sobre milhões de filmes, o que o torna um recurso valioso para analisar dados e criar modelos de previsão. Com acesso a estas informações, podemos explorar a relação entre as diferentes características de um filme e a sua aceitação pelo público.

1.2- Objetivos

Neste trabalho explorei diversas técnicas de *machine learning* para construir um modelo capaz de prever o sucesso de um filme com base em atributos específicos. Portanto neste projeto irei tentar perceber que fatores levam a audiência a gostar de um filme e quais destes têm um maior impacto na avaliação do mesmo, tendo este TFM como objetivos analisar as diversas variáveis de cada filme e perceber o seu peso explicativo na boa ou má avaliação do mesmo, tentando após isso criar um algoritmo através de processos de *machine learning* que permita prever, com alguma precisão, o sucesso de um determinado filme. Para atingir estes objetivos, explorei o pré-processamento de dados, a seleção de características e técnicas de modelação preditiva. Além disso, avaliarei o desempenho do modelo resultante e interpretarei os principais resultados para fornecer informações significativas para a indústria cinematográfica.

Os dados usados neste projeto são provenientes do IMDb, abrangendo 11 variáveis para quase 1 milhão de filmes e 5 variáveis relativos a mais de 12 milhões de participantes.

1.3- Estrutura

A estrutura deste projeto está dividida em 6 capítulos, de forma a estruturar e organizar, da melhor forma, a informação apresentada.

O primeiro capítulo oferece uma introdução ao tema, é apresentado o enquadramento do projeto, assim como os seus os objetivos gerais e a estrutura subsequente.

No segundo capítulo realiza-se uma análise da revisão de literatura relacionada ao campo de estudo. Este capítulo aborda trabalhos anteriores de diversos autores e a sua eficácia em prever o sucesso de filmes. Esta revisão de literatura serve como alicerce para a abordagem metodológica adotada na pesquisa.

No terceiro capítulo descrevemos em detalhe a metodologia utilizada para atingir com objetivos estabelecidos. Isso incluí a seleção da abordagem metodológica e os seus passos.

No quarto capítulo são apresentados os resultados, seguindo as fases da metodologia CRISP-DM. Esta estrutura permite uma análise completa e detalhada dos resultados obtidos.

No quinto capítulo é feita uma discussão entre os resultados alcançados, relacionando-os com os trabalhos analisados na revisão de literatura.

Por fim, o sexto capítulo engloba as conclusões principais do estudo, destacando as suas contribuições para o campo e identificando eventuais limitações. Além disso, são apresentadas sugestões para trabalhos futuros, indicando direções que podem ser exploradas para ampliar o conhecimento neste domínio.

Concluindo, este trabalho final de mestrado representa uma exploração significativa na interseção entre a análise de dados, *machine learning* e a indústria do cinema, com o objetivo final de melhorar a capacidade de prever o sucesso de um filme, proporcionando um benefício tangível para a indústria e, em última análise, para os amantes de cinema em todo o mundo.

Capítulo 2: Revisão de literatura

O que torna um filme bem-sucedido podem ser diversos fatores, dependendo do ponto de vista da pessoa que o está a analisar. Durante a revisão de literatura foram identificados vários artigos que definiam um filme bem-sucedido de várias formas, para alguns autores o sucesso não era mais do que as receitas da bilheteira, para outros o sucesso era medido pelo lucro que o filme gerava, alguns autores definiam por filme bem-sucedido aquele que tinha uma boa avaliação do público, enquanto outros mediam o sucesso de um filme pela influência que este tinha ao longo do tempo. No contexto deste trabalho, foi decidido definir o sucesso do filme pela forma como este é avaliado pelo público, usando assim as avaliações do filme como medição do nível de sucesso do mesmo.

Outro aspeto relevante consiste na utilização de técnicas de *machine learning.* Com efeito, a sua utilização em várias áreas de conhecimento é referida por diversos autores (referido nomeadamente em Aparicio et,al., 2019, Costa & Aparicio, 2023 ou Arriaga & Costa, 2023).

Podem ser encontrados vários estudos relacionados com a previsão de filmes, tendo sido usadas as mais diversas metodologias. Diferentes estudos apresentaram diferentes precisões e diferentes técnicas de implementação de modelos. Na tabela 1 são mostrados vários estudos relacionados com a previsão de sucesso de um filme.

Gupta et al. (2023) previram o sucesso de um filme usando algoritmos de “Ensemble learning”, estes fundamentalmente seguem uma premissa básica, combinam as decisões de múltiplos modelos para melhorar a performance geral (Dietterichl, 2002), o modelo em que obtiveram uma melhor precisão foi o “Gradient Boosting”. Neste artigo foram explorados dados abrangentes, incluindo informações sobre o elenco, diretor, género, orçamento e avaliações iniciais do filme. O seu *dataset* final foi de 1951 filmes.

Dhir & Raj (2018) usaram o algoritmo de aprendizagem automática Random Forest para prever o sucesso de bilheteira dos filmes, obtendo uma precisão de 61%. Estes autores usaram APIs para extrair informações de várias redes sociais, como o Twitter e o Youtube, e analisar o sentimento dos interessados no filme, a base de dados destes autores era composta por 28 variáveis relativas a 5043 filmes. Quader et al. (2017) também tinham como objetivo analisar o sucesso de bilheteira, para tal juntaram dados de várias fontes para analisar os atributos e prever o sucesso dos filmes. Foram usados dois modelos, um destes utilizou apenas atributos antes da estreia do filme nas salas de cinema, e o outro todos os atributos. Os seus melhores resultados foram obtidos usando uma rede neuronal onde conseguiram uma precisão de 89.27%. Jain (2013) empregou uma metodologia completamente diferente, olhando para o aspeto das redes sociais, nomeadamente o Twitter, para prever o sucesso de um filme. Através de um API, o autor coligiu, aleatoriamente, informações relativas a 200 "tweets" para cada filme. Foram escolhidos 24 filmes de 2009 para serem base de treino e 6 filmes para serem o conjunto de teste. O autor criou ainda uma métrica que apelidou de PT-NT ratio definido por ser um rácio dos "tweets positivos" (Positive tweets) a dividir pelos "tweets" negativos (Negative tweets). Este PT-NT ratio foi útil para definir o sucesso de um filme e para avaliar o sentimento do público, obtendo uma previsão comparativamente baixa em relação aos outros autores de 64%.

Autores	**Objetivo da previsão**	**Algoritmo usado**	**Precisão (%)**
Gupta et al., (2023)	*Box office*	*Gradient Boosting*	84.13
Dhir & Raj, (2018)	*Rating*	*Random Forest*	61.00
Quader et al., (2017)	*Box office*	*MLP*	89.27
Jain (2013)	*Box office*	*PT-NT Ratio*	64.40
Mhowwala et al., (2020)	*Rating*	*XGBoost*	95.30
Hsu et al. (2014)	*Rating*	*Multiple Linear regressions*	81.86
Çizmeci & Ögüdücü (2018)	*Rating*	*Factorization Machines*	88.00
Bristi et al. (2019)	*Rating*	*Random Forest*	98.73
Abidi et al. (2020)	*Rating*	*GLM*	76.60

Tabela 1- Modelos de previsão do sucesso de filmes: Objetivos, Algoritmos e Precisões

Mhowwala et al. (2020), por sua vez, usaram dados característicos dos filme e dados de redes sociais, obtidos de várias fontes e APIs, para estimar um modelo. Foram combinados atributos muito interessantes para a análise como a contagem de visitas da página da Wikipédia, o número de comentários no trailer do Youtube, assim como o número de gostos e de visualizações e foi ainda feita uma análise de sentimento nos comentários do vídeo. Os autores usaram um modelo XGBoost, que é uma otimização do modelo Gradient Boosting para prever o sucesso, obtendo uma precisão bastante boa de 95%.

Em Hsu et al. (2014), foi também previsto o sucesso de filmes através do rating do público. Estes usaram uma combinação de modelos lineares para chegar a uma precisão de 81%.

Çizmeci & Ögüdücü, (2018) usaram uma abordagem de *Factorization Machines* para prever a classificação de filmes, combinando dados de filmes e características de redes sociais. Foram usados e combinados dados de diferentes fontes para criar uma base de dados personalizada que contém filmes lançados no ano 2017 nos Estados Unidos.

Bristi et al., (2019) desenvolveram um modelo com dados da Wikipédia e do IMDb para prever as avaliações dos utilizadores de sites de análise de filmes como Rotten Tomato ou o Metacritic. Foram extraídos cerca de 250 filmes do ano de 2018 para realizar o treino do modelo.

No trabalho realizado por Abidi et al., (2020) foi usado um modelo linear generalizado (GLM) que é uma evolução dos modelos lineares convencionais. A precisão deste modelo foi de aproximadamente 77% e a variável que, segundo o autor, mais influencia o sucesso de um filme é a avaliação do diretor.

Estes estudos representam uma variedade de abordagens algorítmicas e metodologias para a previsão de sucesso de filmes, com variadas precisões.

Capítulo 3: Metodologia

Foram identificadas diversas abordagens metodológicas, referenciadas na literatura (Costa & Aparicio, 2020, 2021). Tendo-se optado pelo CRISP-DM (CRoss Industry Standard Process for Data Mining), que foi escolhida devido à sua ampla aplicabilidade. O CRISP-DM foi desenvolvido em 1996 por quatro líderes do mercado em crescimento de extração de dados (Daimler-Benz, IntegralSolutions Ltd. (ISL), NCR, e OHRA). O CRISP-DM é um guia completo para a realização de projetos de extração de dados, adequado para indivíduos de todos os níveis de experiência, desde principiantes a especialistas (Costa & Aparicio, 2020, 2021).. Proporcionando um processo estruturado para a recolha de dados e o desenvolvimento de modelos de uma forma acessível e fácil de compreender.

O CRISP-DM decompõe o ciclo de vida de um projeto de extração de dados em seis fases:

1. Identificação das necessidades da indústria *(Business Understanding)* – O que é que a indústria precisa?
2. Compreensão dos dados *(Data Understanding)* – De que dados dispomos/necessitamos? Estão limpos?
3. Preparação dos dados *(Data Preparation)* – Como é que organizamos os dados para modelação?
4. Modelação *(Modeling)* – Que técnicas de modelação devemos aplicar?
5. Avaliação *(Evaluation)* – Qual o modelo que melhor se adequa aos objetivos do negócio?
6. Implementação *(Deployment)* – Como é que as partes interessadas poderão usufruir dos resultados?

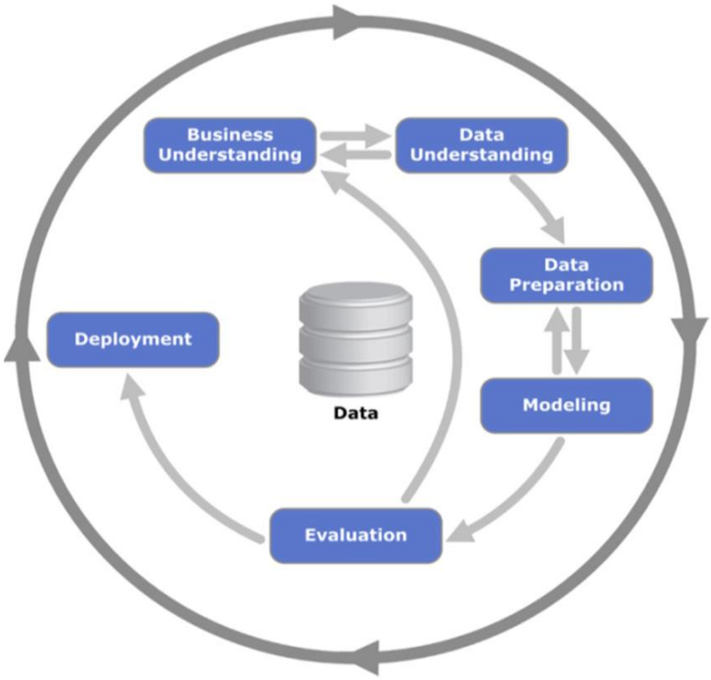

Figura 1- Abordagem CRISP-DM (Chapman et al., 1999)

Na primeira fase, conhecida por ser a base do processo, são determinados os objetivos do projeto, avalia-se a situação, os requisitos do projeto e os recursos disponíveis. Além dos benefícios comerciais, a metodologia CRISP-DM também se foca no sucesso do ponto de vista da modelação dos dados. Após definidos os objetivos, elabora-se o plano do projeto.

A fase seguinte é a da compreensão dos dados, que tem como função identificar, coletar e analisar os *datasets* que irão ajudar a alcançar os objetivos do projeto.

Segue-se então a terceira fase, a preparação dos dados, composta por algumas etapas principais: a seleção dos dados que irão ser usados, incluindo os *datasets* relevantes e excluindo os restantes, a limpeza dos mesmos, a integração dos dados entre as diversas fontes e a transformação da informação conforme necessário. Esta fase é das fases mais importantes e trabalhosas do projeto.

Na fase da modelação dos dados são construídos vários modelos com diversas técnicas de modelação diferentes. Nesta fase deve selecionar-se que algoritmos devem ser usados e dividir os dados em treino e teste. Após tudo isso serão criados os modelos que irão ser analisados e comparados entre si, escolhendo aqueles que melhor satisfaçam os critérios de sucesso estabelecidos.

Na quinta fase é feita uma avaliação mais abrangente dos resultados alcançados. Verifica-se se os resultados obtidos são aplicáveis ao negócio, revê-se o trabalho realizado para tentar perceber se alguma informação foi negligenciada e se todos os passos foram corretamente aplicados. Por fim determinam-se os próximos passos, se devemos avançar para a implementação dos resultados ou se devemos voltar ao princípio e usar a informação obtida para iniciar um novo projeto. Caso se conclua que se deve avançar, chega-se por fim à sexta e última fase do projeto, onde se cria um plano para a implementação dos resultados, verificam-se os resultados obtidos com a implementação e por fim faz-se uma retrospetiva sobre o que correu bem e o que podia ter sido melhor. De referir que esta abordagem pode ainda ser englobada numa perspetiva de *design science* (Aparicio, et al. 2023).

Na elaboração do projeto foi utilizada a linguagem de programação Python através do software Google Colab. O uso do Python deve-se ao facto de esta ser uma linguagem de programação amplamente reconhecida pela sua versatilidade e facilidade de aprendizagem. O Google Colab é uma ferramenta da Google, que permite escrever e executar código através do navegador de internet, sendo especialmente focado para análise de dados e *machine learning*. Um dos pontos positivos desta ferramenta é que este é hospedado nos servidores da Google, não necessitando de nenhuma configuração para usar e permitindo que o código seja executado através de vários dispositivos que de outra forma não o conseguiriam correr, como por exemplo através do telemóvel.

A combinação de uma linguagem acessível e poderosa com a flexibilidade de trabalhar num ambiente de nuvem facilitou a realização das análises complexas e a implementação de soluções de forma eficaz.

Capítulo 4: Resultados

4.1- Identificação das necessidades da indústria *(Business Understanding)*

O setor de entretenimento, particularmente a indústria cinematográfica, é um dos mercados mais competitivos e voláteis da atualidade. Com milhões de euros investidos em cada produção cinematográfica, é imperativo para os estúdios e investidores minimizar os riscos e maximizar o potencial retorno d o investimento. Neste contexto, a análise de dados e a aplicação de algoritmos de *machine learning* tem o potencial de fornecer *insights* valiosos para tomar decisões informadas sobre a produção e o lançamento de filmes. O objetivo deste trabalho final de mestrado é aplicar técnicas de *data mining* e *machine learning* para desenvolver um modelo preditivo capaz de prever o sucesso de um filme. O sucesso de um filme pode ser definido de várias formas, incluindo a receita de bilheteira, a receção da crítica e a popularidade entre o público-alvo. O foco principal deste projeto será a previsão da popularidade entre o público-alvo, mas os conhecimentos adquiridos podem ser aplicados a outras formas de sucesso cinematográfico.

A indústria cinematográfica enfrenta desafios significativos, incluindo o alto custo de produção, a concorrência intensa e a crescente importância da análise de dados na tomada de decisões. A capacidade de prever o sucesso potencial de um filme antes da sua produção e lançamento pode reduzir os riscos financeiros e aumentar a probabilidade de retorno sobre o investimento. Além disso, essa análise pode beneficiar os estúdios, produtores e investidores, ajudando-os a alocar os recursos disponíveis de maneira mais eficaz.

Para atingir os objetivos do projeto, responde-se às seguintes questões:

- Quais são os principais fatores que influenciam o sucesso de um filme?
- Que variáveis são mais relevantes para prever a avaliação do público?
- Que algoritmos de *machine learning* são mais adequados para criar um modelo de previsão do sucesso de um filme?
- Qual é a precisão do modelo na previsão do sucesso dos filmes com base nos dados disponíveis?

A conclusão bem-sucedida deste projeto poderá ter um impacto significativo na indústria cinematográfica, proporcionando uma ferramenta poderosa para a tomada de decisões. Os benefícios esperados incluem a redução de riscos financeiros, a otimização dos recursos, o aumento da eficácia das estratégias de marketing e, em último caso, a maximização do retorno no investimento feito na produção cinematográfica.

4.2- Compreensão dos dados *(Data Understanding)*

A segunda fase do processo trata da recolha e análise dos dados. A qualidade e a relevância dos dados são fundamentais para o sucesso de qualquer análise de *machine learning*. Os dados utilizados neste estudo foram obtidos a partir do IMDb (Internet Movie DataBase), uma das fontes de informação mais completas e fiáveis sobre filmes e séries de televisão. O IMDb disponibiliza vários conjuntos de dados acessíveis ao público que contêm informações detalhadas sobre filmes, incluindo dados relacionados com a equipa de produção, elenco, classificação dos espetadores, as datas de lançamento e muito mais.

Entre os vários conjuntos de dados disponíveis no IMDb, foram selecionados quatro como os mais relevantes para o presente projeto:

Atributo	Descrição
`tconst`	ID do filme
`titleType`	Formato (ex. filme, curta, série de tv, etc)
`primaryTitle`	Título mais popular
`originalTitle`	Título original, na língua de origem
`isAdult`	Se o filme é adulto ou não
`startYear`	Ano de lançamento
`endYear`	Em caso de série de TV ano de fim
`runtimeMinutes`	Duração em minutos
`genres`	Inclui até 3 géneros associados ao filme

Tabela 2- Atributos e descrição dos fiacheiro title.basics.tsv.gz (fonte: https://datasets.imdbws.com/title.basics.tsv.gz)

Atributo	Descrição
`nconst`	ID do participante
`primaryName`	Nome do participante
`birthYear`	Ano de nascimento
`deathYear`	Ano de morte, se aplicável
`primaryProfession`	Top 3 profissões do participante
`knownForTitles`	Filmes pelo qual é conhecido

Tabela 3- Atributos e descrição do ficheiro name.basics.tsv.gz (Fonte: https://datasets.imdbws.com/name.basics.tsv.gz)

Atributo	Descrição
`tconst`	ID do filme
`averageRating`	Média das classificações individuais
`numVotes`	Número de votos que o filme recebeu

Tabela 4- Atributos e descrição do ficheiro title.ratings.tsv.gz (Fonte: https://datasets.imdbws.com/title.ratings.tsv.gz)

Atributo	Descrição
`tconst`	ID do filme
`directors`	Diretores do filme
`writers`	Guionistas do filme

Tabela 5- Atributos e descrição do ficheiro title.crew.tsv.gz (Fonte: https://datasets.imdbws.com/title.crew.tsv.gz)

Os dados recolhidos no IMDb são variados e ricos em informação. No entanto é importante notar que os dados podem conter informações em falta ou duplicadas, o que exigirá um processo de limpeza e preparação dos dados antes da análise.

É fundamental reconhecer que a qualidade e a atualidade dos dados são de extrema importância para a precisão da análise e modelação. A qualidade dos dados é um fator determinante para a eficácia deste estudo. Além disso a atualidade dos dados é essencial para refletir as condições atuais da indústria cinematográfica. Por esse motivo, para garantir que os resultados são consistentes e representativos das últimas tendências, optou-se por utilizar diretamente a base de dados disponível no *site* do IMDb, que é continuamente atualizada pela plataforma. Consequentemente, esta abordagem garante que se está sempre a trabalhar com as informações mais recentes disponíveis no IMDb, mantendo as análises e modelos em linha com as mudanças na indústria cinematográfica.

4.3- Preparação dos dados *(Data Preparation)*

Na fase da “Data Preparation” procedeu-se à limpeza e tratamento dos dados. Antes de iniciar a análise e a modelação, os dados brutos obtidos a partir das fontes do IMDb passaram por um processo de limpeza pormenorizado. No *dataset* referente ao ficheirotitle.basics.tsv.gz na coluna referente ao “`titleType`” estavam contidos vários formatos, todos os dados que não eram “`movie`” ou “`tvMovie`” foram removidos.

```
titleType
movie            659018
short            955021
tvEpisode       7788984
tvMiniSeries      50686
tvMovie          143183
tvPilot               1
tvSeries         250203
tvShort           10033
tvSpecial         43636
video            281252
videoGame         36102
```

Figura 2- Formatos contidos no "titleType" antes da limpeza

```
titleType
movie      659018
tvMovie    143183
```

Figura 3- Formatos contidos no "titleType" após limpeza

No *dataset* incluído no ficheiro title.ratings.tsv.gz foram eliminados todos os filmes que apresentavam uma contagem de votos "`numVotes`" inferior a mil, esta decisão foi tomada para evitar filmes com valores inflacionados ou deflacionados devido ao pouco número de votos.

Os valores em falta de todas as tabelas foram identificados e tratados, no *dataset* title.basics.tsv.gz, a coluna "`endYear`" foi removida devido ao facto de maior parte dos valores desta coluna serem "/NA".

No *dataset* name.basics.tsv.gz foram removidas as colunas "`birthYear` "e "`deathYear`" visto não se considerar relevante o ano de nascimento ou o ano de morte dos participantes para a avaliação de um filme.

Os dados foram também padronizados quando necessário, garantindo que os formatos dos dados fossem consistentes.

Para melhorar a capacidade preditiva do sucesso dos filmes, foram criadas variáveis a partir dos dados originais, tais como a avaliação média do guionista "`avaliação_writer`" e a avaliação média do diretor "`avaliação_directors`", estas foram criadas fazendo a média das avaliações do filme em que o interveniente entrava como diretor ou como guionista.

Foi inicialmente feita uma análise ao número de lançamentos de filme por ano, esta análise permite entender evolução da indústria cinematográfica. Como se pode ver na figura 4, assiste-se a um aumento gradual e constante dos lançamentos de filmes ao longo do tempo. Esta tendência manteve-se até 2020, ano em que a indústria cinematográfica foi significativamente afetada pela pandemia. Nesse ano, registou-se uma queda notável no número de filmes lançados. A partir de 2021, a indústria experimentou um processo de recuperação, embora o número de lançamentos ainda não tenha atingido os níveis anteriores à pandemia. É de notar que os números relativos a 2023 acabam por ser imprecisos, uma vez que o ano ainda não acabou e ainda poderão ser lançados vários filmes até ao fim do mesmo.

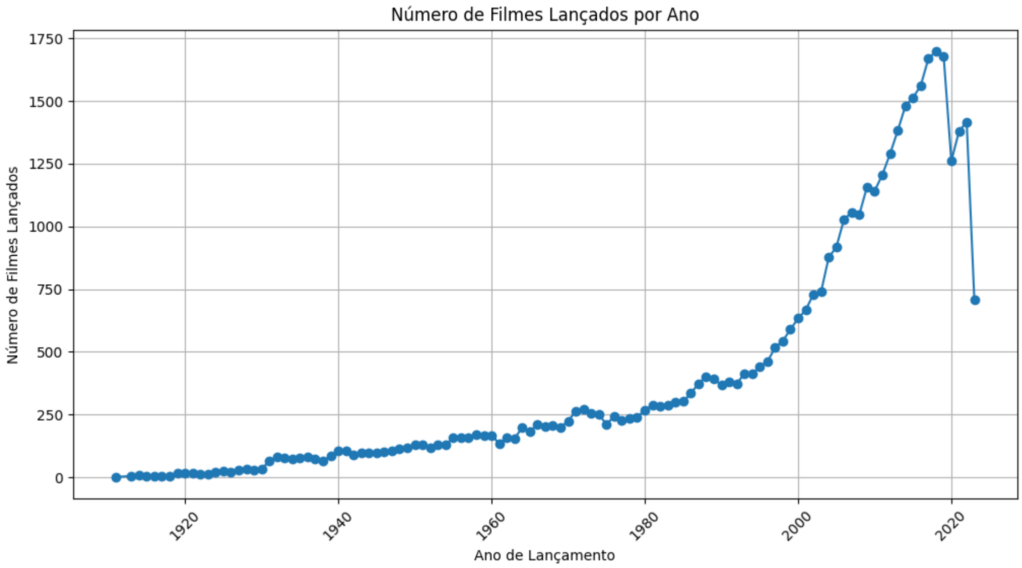

Figura 4- Número de lançamento de novos filmes por ano

Seguidamente, foi feita uma análise da forma como estão distribuídas as avaliações dos filmes (Figura 5 e Figura 6). Esta análise revela informações valiosas sobre a forma como o público avalia e perceciona as produções cinematográficas. As classificações variam numa escala de 1 a 10, em que 1 representa a classificação mais baixa e 10 a mais alta. Para uma análise mais pormenorizada, as classificações foram agrupadas em categorias com base no número de classificações recebidas por filme, resultando nos grupos presentes na tabela 6, a observação da distribuição revela que a maioria dos filmes se concentra nas gamas de classificação superiores, com um pico notável no grupo (6, 7]. Esta concentração sugere que muitos filmes são amplamente apreciados pelo público, com classificações que variam entre 6 e 7. No entanto, é também importante notar que existe uma presença considerável de filmes nos grupos (4, 5] e (7, 8], o que indica uma variedade de perceções do público sobre as produções cinematográficas.

Os grupos de classificação mais extremos, ou seja, os que têm classificações muito baixas (grupo (0, 1]) e muito altas (grupo (9, 10]), representam uma minoria. O grupo (0, 1] contém apenas cinco filmes, o que sugere que um número muito limitado de filmes recebeu classificações extremamente negativas. Por outro lado, o grupo (9, 10] inclui apenas 31 filmes, indicando que apenas um grupo selecionado de filmes recebeu classificações excecionalmente positivas. Esta distribuição de classificações é fundamental para a investigação, uma vez que ajuda a compreender a amplitude das perceções do público relativamente ao sucesso dos filmes e a identificar padrões nas classificações que podem ser utilizados nos nossos modelos de previsão.

Grupo	Frequência
(0,1]	5
(1,2]	140
(2,3]	547
(3,4]	1654
(4,5]	3980
(5,6]	9475
(6,7]	16485
(7,8]	10029
(8,9]	1294
(9,10]	31

Tabela 6- Frequência de classificação por grupo

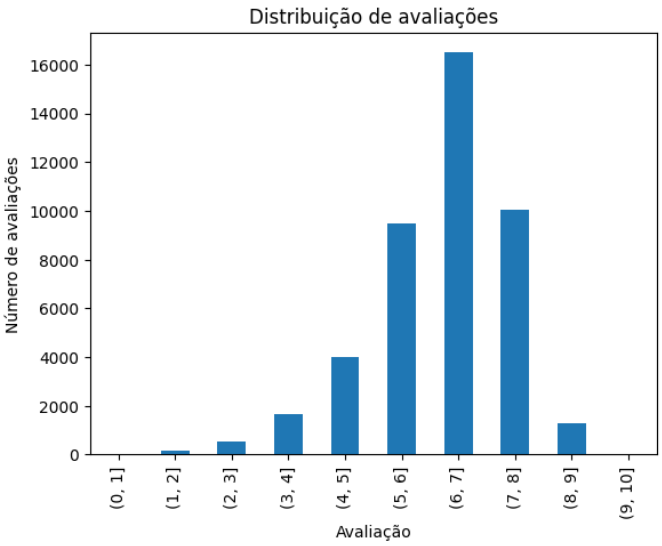

Figura 5- Distribuição de avaliações

Foi feita uma análise de como a duração dos filmes impactava a sua classificação, para tal, foram divididos os filmes em intervalos de 60 minutos. Como é possível observar na figura 6, é possível retirar que há uma pequena tendência de crescimento da média das avaliações ao aumentar a duração dos filmes. É de notar que no mesmo gráfico não existe representação de filmes entre 600 e 720 minutos pois não há nenhum filme com estas características.

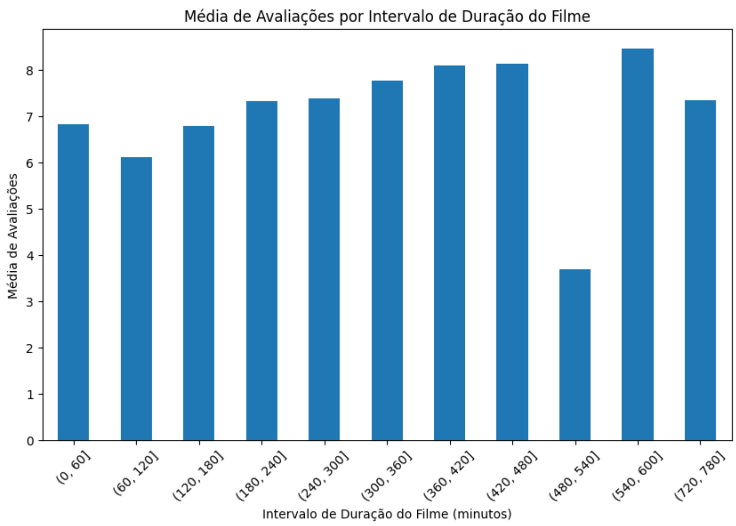

Figura 6- Média das avaliações por intervalo temporal

Intervalo	Avaliação Média	Contagem
(0, 60]	6.83	431
(60, 120]	6.12	35549
(120, 180]	6.78	7307
(180, 240]	7.33	274
(240, 300]	7.37	35
(300, 360]	7.77	10
(360, 420]	8.10	3
(420, 480]	8.13	3
(480, 540]	3.70	1
(540, 600]	8.45	2
(720, 780]	7.35	2

Tabela 7- Média das avaliações por intervalo temporal

No entanto, a duração dos filmes não afeta os diferentes géneros da mesma forma, na figura 7 e 8 conseguimos comparar a relação das avaliações e da duração entre os filmes com o género comédia e o género ação, sendo observável que nos filmes de comédia o aumento da duração do filme tem uma correlação positiva com o aumento das avaliações enquanto nos filmes de ação não se sente tanto essa tendência.

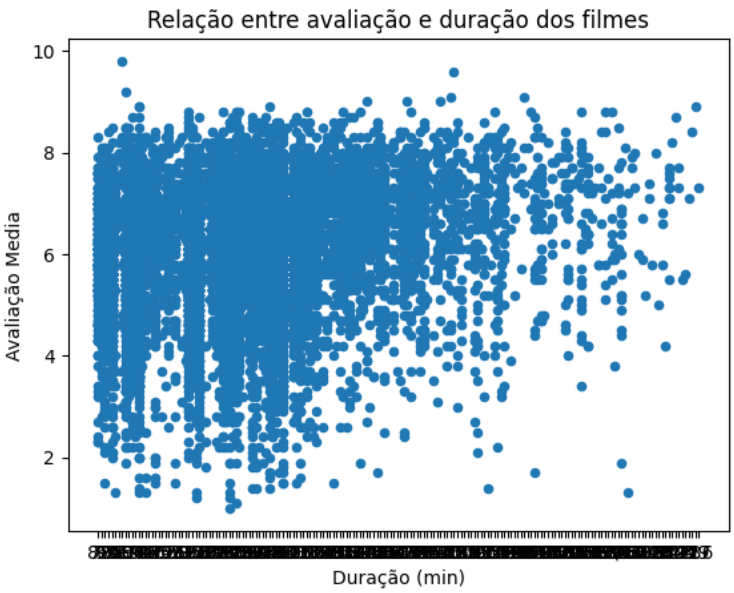

Figura 7- Relação entre avaliação e duração em filmes de comédia

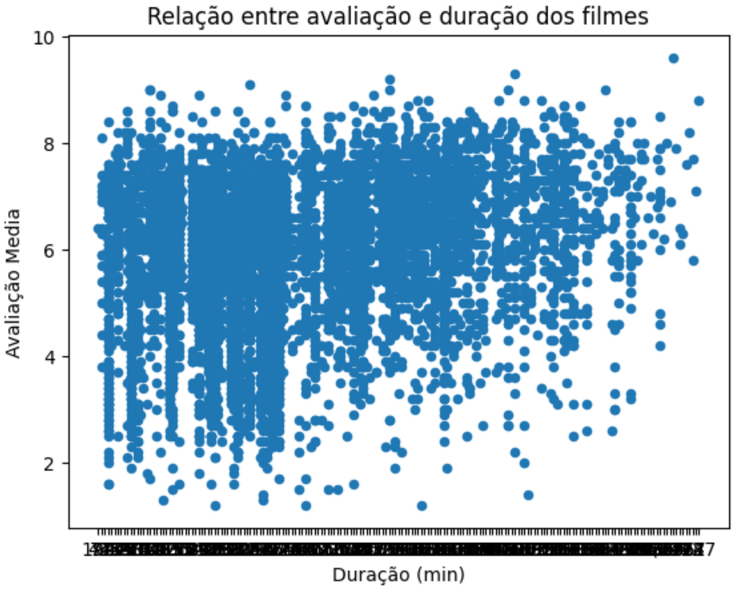

Figura 8- Relação entre avaliação e duração em filmes de ação

Foi também feita uma análise sentimento dos títulos dos filmes usando a biblioteca Vader (*Valence Aware Dictionary and sEntiment Reasoner*), que é uma ferramenta amplamente reconhecida para a análise de sentimentos em textos de linguagem natural, composta por uma coleção de palavras com pontuações de sentimento predefinidas (Hutto & Gilbert, 2014), tendo sido aplicada em vários domínios (e.g. Aparicio et al., 2021). Esta análise de sentimentos tem como objetivo atribuir uma pontuação de sentimentos a cada título de filme, variando entre -1 (negativo) e 1 (positivo). A razão para realizar esta análise de sentimento é explorar a possibilidade de que a perceção inicial do público em relação a um filme pode ser influenciada pelo título. Por conseguinte, ao incluir esta pontuação de sentimento como uma variável no conjunto de dados, pretendemos avaliar se os títulos com um sentimento mais positivo ou negativo têm uma correlação com o êxito de um filme. Isto pode ajudar a compreender se a escolha de um título cativante e emocionalmente envolvente desempenha um papel importante no sucesso de um filme. Após a análise de sentimento, foi gerada uma representação visual das palavras mais usadas nos títulos dos filmes por meio de uma *wordcloud*. O processo de criação da *wordcloud* presente na figura 9 envolveu a remoção de *stopwords* (palavras de uso comum que não agregam significado) e a contagem da frequência de cada palavra.

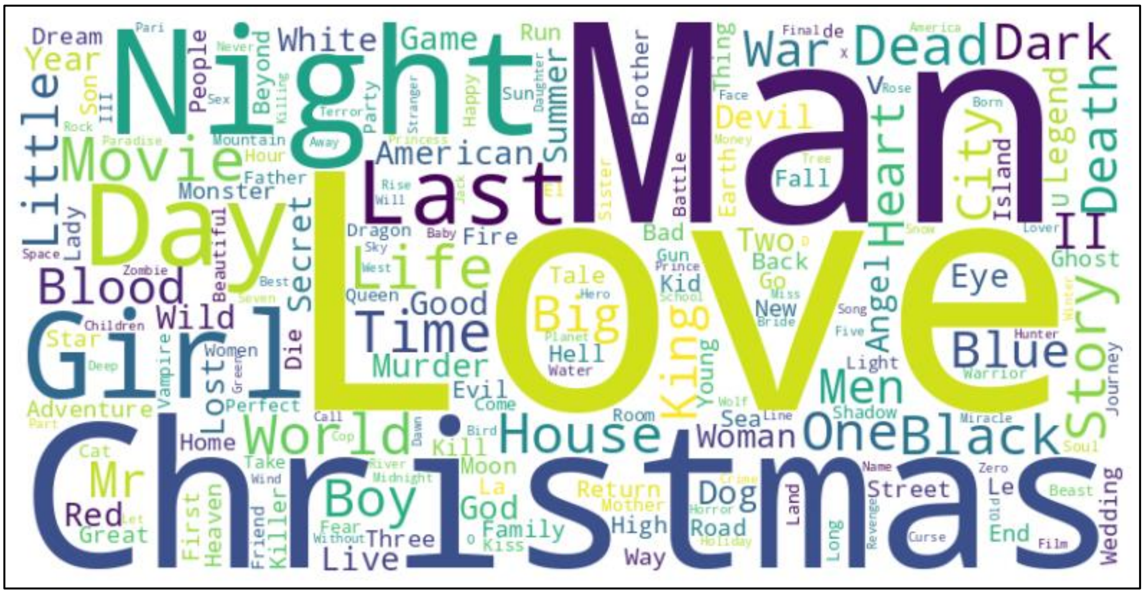

Figura 9- Wordcloud das palavras mais usadas nos títulos de filmes

A representação gráfica resultante destaca as palavras que ocorrem com maior frequência, fornecendo informações valiosas sobre as temáticas enfatizadas nos filmes da amostra. Ao examinar a nuvem de palavras, observámos palavras proeminentes que podem influenciar a perceção que o público tem de um filme. Nomeadamente, palavras como "Love", "Man" ou "Night" surgiram como termos frequentes nos títulos. Para uma compreensão mais detalhada da distribuição das palavras mais utilizadas, realizamos também uma análise de frequência das palavras, apresentada no gráfico de barras na figura 10. Este gráfico revela que a palavra “Love” é significativamente mais frequente do que qualquer outra palavra, chegando a ter o dobro da utilização que qualquer outra palavra que não esteja no “Top 3” de utilização.

Estas análises fornecem informações valiosas sobre as tendências linguísticas dos títulos de filmes, permitindo avaliar o impacto potencial das escolhas linguísticas na perceção e no sucesso dos filmes por parte do público.

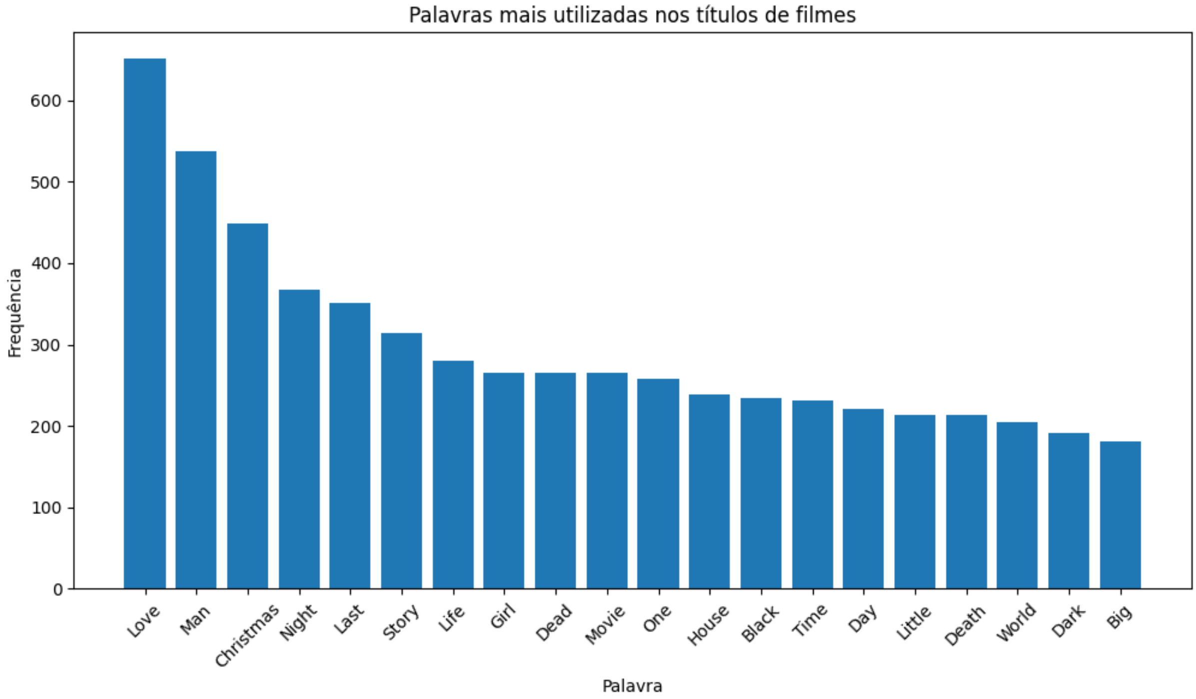

Figura 10- Gráfico de barras com frequência de cada palavra

Além disso, foi criada uma variável para contar o número de caracteres do título "`Title_Char`", servindo para analisar se um filme deveria ter um título maior ou menor de forma a ser mais atrativo.

Durante a preparação dos dados, foram agregadas informações de várias fontes. Inicialmente, combinámos os dados dos conjuntos "title.basics.tsv.gz" e "title.ratings.tsv.gz". Esta fusão permitiu-nos acrescentar informações valiosas, como a classificação média do filme e o número de votos, ao conjunto de dados principal que contém informações básicas sobre cada filme. Este novo conjunto de dados resultou na criação de um *dataframe* denominado "Movies_rating". Além disso, o *dataframe* "Movies_rating" foi enriquecido com informações sobre a equipa de produção, incluindo os nomes dos argumentistas e realizadores, a partir do arquivo "title.crew.tsv.gz". Isto foi feito para compreender melhor a contribuição da equipa de produção para o sucesso de um filme. O resultado foi um dataframe chamado "Movies_crew".

O dataframe "Movie" foi, então, criado através da combinação de "Movies_crew" com as variáveis criadas durante a preparação dos dados, tais como " `avaliação_writer` ", " `avaliação_directors` ", " `sentimento` " e "`Title_Char`". Este passo de agregação permitiu consolidar toda a informação relevante num único conjunto de dados, que servirá de base para a modelação de *machine learning*.

Variável	Média	Mediana	Mínimo	Máximo	Desvio Padrão
averageRating	6.25	6.40	1.00	9.90	1.18
runtimeMinutes	104.30	100.00	11.00	776.00	23.60
startYear	1999	2007	1911	2023	22
isAdult	0.00	0.00	0.00	1.00	0.02
numVotes	24135	3177	1000	2805376	90504
avaliação_writer	6.25	6.40	1.00	9.90	1.14
avaliação_directors	6.25	6.43	1.00	9.90	1.03
sentimento	-0.01	0.00	-0.94	0.91	0.24
Title_Char	15.66	14.00	1.00	104.00	8.63

Tabela 8- Visão geral das variáveis dos filmes

Após a compilação dos dados e a criação do dataframe " Movie", foi feita uma seleção cuidadosa das variáveis para determinar quais as características a incluir na modelação.

Este processo de seleção envolveu várias etapas, entre as quais, a análise de correlação, sendo que as correlações entre as variáveis foram analisadas para identificar relações lineares entre elas. As variáveis altamente correlacionadas foram avaliadas para evitar a multicolinearidade e reduzir o número de características redundantes. As variáveis foram avaliadas em termos da sua importância para o objetivo do projeto, ou seja, prever o sucesso de um filme. As que não contribuíam significativamente para a capacidade de previsão foram excluídos.

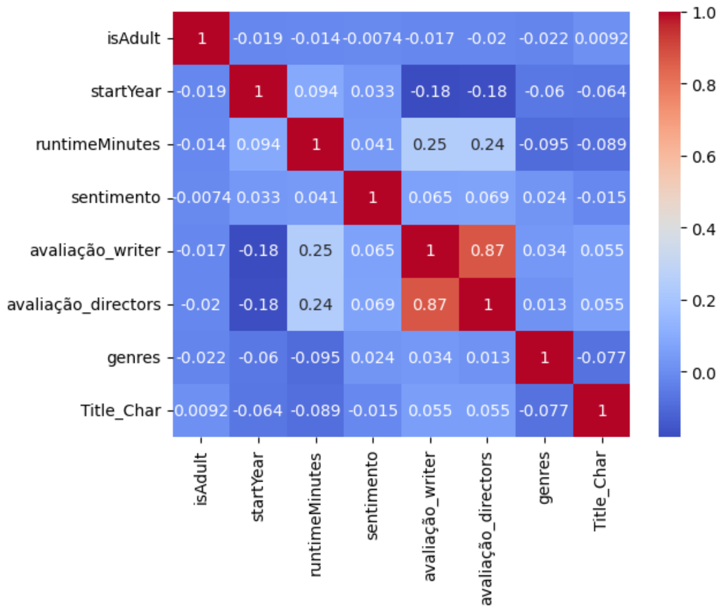

Figura 11- Matriz de correlação das variáveis

Como é possível observar na figura 11, temos duas variáveis altamente correlacionadas, a "`avaliação_directors`" e a "`avaliação_writer`". Foram então calculados os Valores de Inflação da Variância (VIF) para cada variável independente para identificar a multicolinearidade.

Variável	VIF
`isAdult`	1.001518
`startYear`	54.599573
`runtimeMinutes`	22.694865
`sentimento`	1.007052
`avaliação_writer`	130.994372
`avaliação_directors`	157.309294
`genres`	4.557174
`Title_Char`	4.383222

Tabela 9- VIF das variáveis do df "Movie"

Devido à grande correlação e aos valores elevados do VIF, procedeu-se à padronização dos dados e à análise de componentes principais ("PCA" *Principal Component Analysis*) de Karl Pearson (Pearson, 1901), esta é uma técnica de redução de dimensionalidade dos dados, que transforma as variáveis originais num conjunto de componentes principais não correlacionados. Foram então criados cinco componentes principais.

	is Adult	start Year	runtime Minutes	Senti mento	avaliação _writer	avaliação _directors	genres	Title _Char
PC1	-0.02402	-0.2015	0.2860	0.0884	0.6585	0.6575	0.0131	0.0548
PC2	-0.0907	0.6048	0.5866	0.2054	-0.0348	-0.0275	-0.2934	-0.3894
PC3	0.2378	0.0502	0.0838	-0.2579	-0.0123	0.0019	-0.7371	0.5691
PC4	0.9496	-0.0689	0.0636	0.0141	0.0058	0.0005	0.0820	-0.2874
PC5	0.1152	0.0948	-0.1651	0.9046	-0.0260	-0.0205	-0.0120	0.3618

Tabela 10- Componentes Principais

Criados os componentes principais foram estudados os seus coeficientes para analisar como estes se relacionavam com as variáveis originais. O primeiro componente principal (PC1) tem coeficientes significativos para as variáveis "`avaliação_writer`" e "`avaliação_directors`". Isto sugere que este componente esteja diretamente relacionado com a avaliação do elenco do filme, sendo que quanto melhor a avaliação média destes, melhor a avaliação do filme. O segundo componente PC2, tem coeficiente significativo para as variáveis "`startYear`" e "`runtimeMinutes`", sugerindo que o componente 2 esteja relacionado com as características temporais do filme. Por sua vez, o PC3 está relacionado positivamente com o número de caracteres do título do filme e negativamente com o género.

O quarto componente principal é explicado quase na sua totalidade pela variável "`isAdult`", implicando uma relação com a natureza de um filme ser para maiores de dezoito ou não. Por fim, o último componente principal PC5, é positivamente relacionado com a análise de sentimento feita aos títulos dos filmes.

A seleção criteriosa das variáveis foi realizada para otimizar o desempenho dos modelos de *machine learning*, mantendo apenas as características mais relevantes e informativas para a previsão do sucesso de um filme. Após a preparação pormenorizada dos dados, o conjunto de dados estava pronto a ser utilizado para construir, treinar e avaliar os modelos de *machine learning*.

4.4- Modelação *(Modeling)*

A fase de modelação desempenha um papel crucial no projeto. Com o propósito de prever o sucesso dos filmes, os dados foram divididos em conjuntos de treino e teste para a construção e validação dos modelos de *machine learning* respetivamente. A proporção de divisão foi cuidadosamente escolhida para garantir uma avaliação adequada do desempenho do modelo, sendo a proporção usada 80% treino e 20% teste. Foi feita uma seleção cuidadosa de modelos de *machine learning* que iriam ser usados na modelação dos dados, tendo em conta as suas diferentes capacidades e propriedades.

Inicialmente foi feita uma análise dos resultados da regressão OLS. O objetivo principal foi verificar se todas as variáveis eram significativas para o modelo e observar a influência positiva ou negativa destas no resultado final da estimação.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          averageRating   R-squared:                       0.857
Model:                            OLS   Adj. R-squared:                  0.857
Method:                 Least Squares   F-statistic:                 5.241e+04
Date:                Sun, 08 Oct 2023   Prob (F-statistic):               0.00
Time:                        13:03:06   Log-Likelihood:                -26627.
No. Observations:               43565   AIC:                         5.327e+04
Df Residuals:                   43559   BIC:                         5.332e+04
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          6.2498      0.002   2925.480      0.000       6.246       6.254
PC1            0.7634      0.001    511.195      0.000       0.760       0.766
PC2           -0.0403      0.002    -20.610      0.000      -0.044      -0.036
PC3           -0.0167      0.002     -8.187      0.000      -0.021      -0.013
PC4            0.0076      0.002      3.555      0.000       0.003       0.012
PC5           -0.0317      0.002    -14.731      0.000      -0.036      -0.027
==============================================================================
Omnibus:                     8101.294   Durbin-Watson:                   1.813
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            46880.351
Skew:                          -0.773   Prob(JB):                         0.00
Kurtosis:                       7.841   Cond. No.                         1.44
==============================================================================
```

Figura 12- Regressão OLS

Apesar do resultado satisfatório do $R^{2 \text{ de } 0.57}$, foram ainda utilizados outros algoritmos, com recurso a aprendizagem automática, para tentar obter um modelo com resultados ainda melhores.

Algoritmo	MSE	RMSE	MAE	R^2
Regressão Linear	0.1928	0.4391	0.3221	0.8597
Regressão de Ridge	0.1928	0.4391	0.3220	0.8597
Regressão de Lasso	0.6742	0.8211	0.6364	0.5094
SVR	0.2277	0.4771	0.3167	0.8343
Random Forest	0.2132	0.4618	0.3418	0.8448
MLP Regressor	0.1687	0.4107	0.3090	0.8772
Gradient Boosting Regressor	0.1659	0.4073	0.2951	0.8793
K-NN	0.1910	0.4371	0.3212	0.8610
TensorFlow	0.0773	0.4371	0.3212	0.8610

Tabela 11- Algoritmos de machine learning e a sua precisão

Como é possível observar na tabela 11, foram selecionados vários algoritmos:

- A **regressão linear** e a **regressão de Ridge**: Ambos os modelos são muito interpretáveis e excelentes para refletir relações lineares entre as variáveis independentes e a variável de dependente. São ideais para situações em que a simplicidade e a interpretabilidade são fatores importantes. Estas regressões demonstraram uma capacidade notável de explicar a variação dos dados, com R^2 de aproximadamente 0.86. Apresentaram, também, valores de erro médio absoluto (MAE) na faixa dos 0.3220 a 0.3221.

- A **regressão de Lasso**: Apesar de ser útil para a escolha de características, uma vez que penaliza alguns coeficientes, tornando alguns atributos mais influentes do que outros, pode não ser a escolha ideal quando todas as variáveis são relevantes, o que acabou por acontecer, sendo este modelo o que apresenta o desempenho inferior em comparação com os restantes, com o R^2 mais baixo de 0.5094.

- **SVR** (*Support Vector Regression*): este modelo é eficaz na captura de relações não lineares. Este modelo exibiu um bom desempenho, com um R^2 de 0.8343 e um MAE de 0.3167.

- ***Random Forest***: o modelo *Random Forest* é um método ensemble que constrói e combina centenas ou milhares de árvores de decisão, treinando cada uma num conjunto ligeiramente diferente de dados. A previsão final é dada pela média da previsão das várias árvores individuais. É capaz de lidar com relações não lineares complexas, é resistente ao *overfitting* e lida bem com *outliers*. Em resumo, o *Random Forest* é uma técnica poderosa, amplamente utilizada para resolver problemas de classificação e regressão. Apresentou resultados sólidos, com R^2 de 0.8448 e um MAE de 0.3418.

- **MLP Regressor** (*Multilayer Perceptron Regressor*): Redes neuronais; como este modelo, são muito flexíveis. Este modelo destacou-se, com um bom desempenho geral, apresentando um R^2 de 0.8772 e um MAE de 0.3090.

- **Gradient Boosting Regressor**: o modelo Gradient Boosting é um modelo que treina gradualmente vários modelos de forma sequencial, melhorando os modelos anteriores. Também exibiu um excelente desempenho, obtendo um R^2 de 0.8793 e um MAE de 0.3090. Este modelo é conhecido por ser eficaz a modelar relações não lineares e interações complexas dos dados.

- **K-NN (K-Nearest Neighbors)**: O K-NN distingue-se por ser adequado para identificar padrões nos dados e é útil em cenários de baixa dimensionalidade. O modelo obteve resultados consistentes, com um R^2 de 0.8610.

- **TensorFlow**: é um modelo de rede neuronal criado pela Google. Como *rede neuronal* é capaz de modelar relações complexas em grandes conjuntos de dados. No entanto acaba por ser computacionalmente intensivo. Os resultados obtidos através deste modelo foram bastante sólidos.

Com base nos resultados obtidos e nas considerações feitas sobre a utilidade de cada modelo foi possível concluir que os modelos mais complexos, como MLP Regressor e Gradient Boosting, superaram os restantes em termos de desempenho da modelação dos dados.

4.5- Avaliação *(Evaluation)*

Nesta fase, realizou-se uma avaliação abrangente do desempenho do modelo de *machine learning* utilizado neste projeto. A fase de Avaliação é essencial para determinar o quão bem o modelo é capaz de cumprir os objetivos e se as previsões são confiáveis o suficiente para serem aplicadas.

As métricas escolhidas para determinar o melhor modelo foram o coeficiente de determinação (R^2) e o erro absoluto médio (MAE). O modelo será tanto melhor quanto maior for o seu R^2 e menor o seu MAE, e, como tal, conseguimos perceber que o desempenho do modelo Gradient Boosting Regressor foi superior aos restantes, sendo o melhor em ambas as métricas. Após toda a análise foi concluído que o melhor modelo a ser usado seria o Gradient Boosting Regressor pois foi este que previu com maior precisão o sucesso de um filme.

A validação cruzada assume um papel crucial na avaliação dos modelos. Foi utilizada a validação cruzada *k-fold* para avaliar a capacidade de generalização do modelo em diferentes conjuntos de dados. Os resultados da validação cruzada confirmaram a consistência do desempenho do modelo em diferentes divisões dos dados de treino e de teste.

```
Resultados das folds (MSE): [0.08466202 0.08352575 0.08854297 0.08262103 0.08849708]
Média dos resultados (MSE): 0.08556977022232207
```

Figura 13- 5 fold cross-validation

O MSE é uma métrica que mede o erro quadrático médio das previsões do modelo em relação aos valores reais, quanto menor o seu valor, melhor o desempenho do modelo, pois significa que as previsões estão mais próximas dos valores reais. Como é possível observar na figura acima, a média do MSE para todas as *folds* foi de aproximadamente 0.0856, o que indica que, em média, as previsões do modelo têm um erro quadrático médio de 0.0856 em relação aos valores reais.

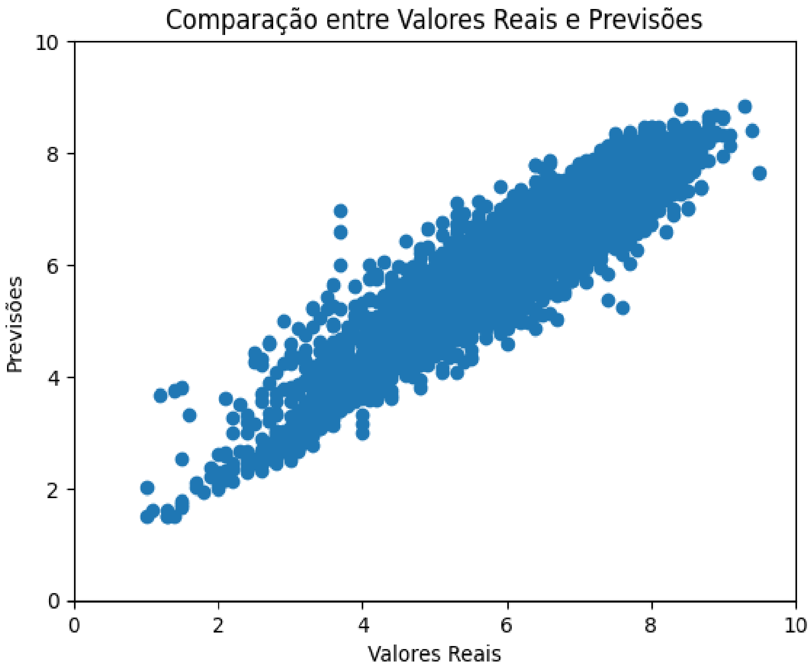

Figura 14- Comparação entre os valores reais e previstos do modelo Gradient Boosting

Na figura 14, conseguimos ver uma figura com a representação visual do desempenho do modelo preditivo. Como é possível observar, há uma precisão bastante grande entre os valores reais e as previsões geradas pelo modelo Gradient Boosting Regressor, ilustrando assim a eficácia do modelo em prever com precisão a relação entre as variáveis independentes e a variável dependente. A proximidade dos resultados indica a capacidade do modelo de minimizar os erros de previsão, resultando num desempenho altamente preciso.

Foram também analisados os resíduos, visto ser uma ferramenta útil para avaliar a qualidade do modelo. Como é facilmente observável na figura 15 os valores residuais estão concentrados perto de zero, o que indica que o modelo está a fazer uma previsão precisa.

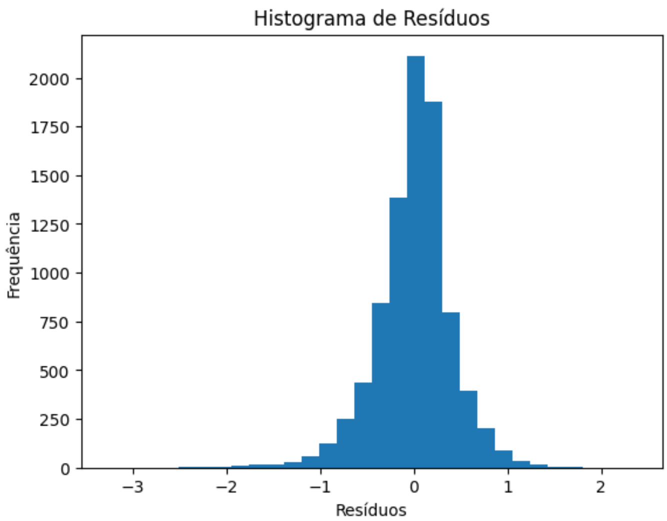

Figura 15- Histograma de Resíduos

4.6- Implementação *(Deployment)*

A fase de Implementação representa a etapa culminante de todo o processo de desenvolvimento de previsão utilizando modelos de aprendizagem automática. Durante esta fase, os modelos desenvolvidos e validados ao longo do projeto serão aplicados num ambiente de produção para que possam ser utilizados de forma prática e benéfica. Esta fase é crucial para transformar os resultados da análise de dados em ações tangíveis e valiosas para a indústria cinematográfica. A implementação deste projeto seria possível através de uma aplicação, onde os *stakeholders* podiam testar as várias combinações de atores, diretores, géneros, etc., de atributos para tentarem obter as melhores combinações possíveis de filme.

A análise SWOT é uma ferramenta de planeamento estratégico muito utilizada que ajuda a avaliar um projeto. O termo "SWOT" é um acrónimo que se refere às quatro dimensões principais que a análise considera: Pontos Fortes (***S**trengths*), Pontos Fracos (***W**eaknesses*), Oportunidades (***O**pportunities*) e Ameaças (***T**hreats*). Cada uma destas dimensões é cuidadosamente examinada para compreender a situação atual e antecipar ações futuras. A análise SWOT, presenta na tabela 12, realça a importância de maximizar os pontos fortes e as oportunidades e de atenuar os pontos fracos e as ameaças durante a fase de implementação. A implementação bem-sucedida de modelos de aprendizagem automática é essencial para obter benefícios reais e mensuráveis na indústria cinematográfica.

Fatores internos	**Forças** -Fácil compreensão da informação apresentada; -Modelos com boa precisão.	**Fraquezas** -Pouca diversidade dos dados; -Possíveis incongruências das fontes de dados; -Complexidade técnica da implantação -Possíveis erros na implementação técnica, o que poderia levar a previsões imprecisas.
Fatores externos	**Oportunidades** -Aplicabilidade a outras áreas de negócio; -Os modelos podem automatizar e melhorar as decisões de investimento; -A eficiência operacional pode ser aumentada, reduzindo custos e riscos financeiros.	**Ameaças** -Crescimento de ferramentas de inteligência artificial concorrentes com maior capacidade de armazenamento; processamento de dados e mais simples na ótica do utilizador.

Tabela 12- Análise swot da solução da previsão utilizando algoritmos de machine learning

Nesta fase, é importante considerar pormenores técnicos, como a escolha da plataforma ou do ambiente de produção onde os modelos serão implementados. Trata-se de descrever o processo de implementação, desde a preparação dos dados de entrada até à configuração dos servidores e à integração com os sistemas existentes na indústria cinematográfica.

O acompanhamento contínuo dos modelos implementados é também fundamental, garantindo que estes mantêm o seu desempenho e a qualidade das previsões ao longo do tempo. Devendo definir-se políticas de adaptação dos modelos para manter a sua eficácia.

A comunicação e a formação são igualmente importantes. A explicação de como os resultados das previsões serão comunicados e utilizados na tomada de decisões é crucial. Se houver necessidade de formação para os utilizadores do sistema, é igualmente importante descrever a forma como essa formação será proporcionada. Além disso, é essencial estabelecer garantias de qualidade e efetuar testes rigorosos antes da implementação total. Isto ajuda a garantir que os modelos irão funcionar como esperado num ambiente de produção.

Deve, por fim, ser elaborado um plano de salvaguarda e de contingência para fazer face a eventuais falhas dos modelos ou à deterioração da qualidade das previsões, garantindo a continuidade das operações.

Capítulo 5: Discussão

No início do projeto, estabelecemos a definição de sucesso de um filme, que pode variar de acordo com diferentes métricas, como as receitas de bilheteira, as críticas e a popularidade entre o público. Embora o nosso foco principal tenha sido a previsão da notoriedade entre o público-alvo, as técnicas e os resultados deste projeto podem ser aplicados a outras definições de sucesso. Ao longo deste estudo, explorámos os fatores que podem ter um impacto significativo na avaliação de um filme por parte do público. Estes fatores incluem a qualidade da equipa de produção, as características temporais, o género, a análise do sentimento nos títulos e a duração dos títulos. Tal como em (Mhowwala et al., 2020) chegámos à conclusão que a avaliação do diretor e do guionista, que por muitas vezes acabam por ser a mesma pessoa, é o atributo mais importante para prever a avaliação do filme.

Foram testados vários algoritmos de aprendizagem automática, desde regressões lineares tradicionais a métodos mais complexos, como *Gradient Boosting* e *Redes Neuronais.* É relevante comparar os resultados com outros estudos e investigações semelhantes. De um modo geral, os nossos resultados estão em linha com as conclusões anteriores na área da análise preditiva do sucesso de filmes. O melhor algoritmo foi o *Gradient Boosting Regressor,* que teve uma precisão de aproximadamente 88%, comparado com outros trabalhos pode dizer-se que é uma precisão alta, havendo poucos estudos com resultados mais elevados. No entanto, na análise feita neste projeto foram considerados quase todos os filmes disponíveis na base de dados do IMDb, sendo no final o modelo estimado para quase 45 mil filmes, o que o diferencia dos trabalhos anteriores onde, na sua maioria, a base de dados era extremamente reduzida e ajustada, havendo trabalhos onde o *dataset* para previsão continha pouco mais de 200 filmes (Bristi et al., 2019).

Capítulo 6: Conclusões, Limitações e Trabalhos Futuros

A indústria cinematográfica é notória pela sua elevada competitividade e volatilidade. Com custos substanciais envolvidos na produção de filmes, a capacidade de minimizar riscos e maximizar retornos é crucial para estúdios e investidores. Neste contexto, a análise de dados e a utilização de algoritmos de aprendizagem automática são ferramentas valiosas para tomar decisões informadas sobre a produção e o lançamento de filmes. Nesta investigação, foram utilizados diferentes algoritmos de aprendizagem automática para construir um modelo capaz de prever o sucesso de um filme. Os modelos baseados nos algoritmos MLP e Gradient Boosting Regressor apresentaram resultados mais consistentes, tendo taxas de sucesso a rondar os 88%. Além destes, também outros algoritmos como o K-NN e a rede neuronal baseada no TensorFlow obtiveram bons resultados.

É importante notar que a nossa investigação tem limitações, como qualquer estudo. Os resultados baseiam-se em dados disponíveis até à data da investigação e a indústria cinematográfica está em constante evolução. Por esse motivo, os resultados podem não captar todas as mudanças e tendências atuais. Houve muitas limitações em termos de dados, por exemplo vários atributos, possivelmente muito relevantes, que não foram facilmente obtidos. Além disso existia a limitação do software, o Google Colab, que apesar de ser uma ferramenta muito útil, tinha várias limitações de processamento na sua versão gratuita.

Para melhorias futuras, mais atributos deveriam ser considerados como mês de lançamento, para estudar como a sazonalidade dos lançamentos impacta a avaliação dos mesmos, o orçamento dos filmes, analisar se os atores são premiados ou não, analisar os principais atores através do número de gostos ou seguidores nas redes sociais, e fazer uma análise de sentimento a comentários nos *trailers* do filme. Além disso, uma análise por género, ao invés de uma análise geral, traria resultados ainda mais consistentes.

A investigação confirma muitas das tendências identificadas na revisão da literatura e destaca a importância da qualidade da equipa de produção, da análise de sentimentos nos títulos e de outros fatores na previsão do êxito dos filmes. Os resultados da investigação também apontam para a eficácia dos algoritmos de aprendizagem automática na previsão do êxito dos filmes. De uma forma geral, esta investigação fornece informações valiosas aos intervenientes na indústria cinematográfica, oferecendo uma abordagem baseada em dados para tomar decisões informadas e reduzir os riscos financeiros na produção e lançamento de filmes.

Referências Bibliográficas

Abidi, S. M. R., Xu, Y., Ni, J., Wang, X., & Zhang, W. (2020). Popularity prediction of movies: From statistical modeling to machine learning techniques. *Multimedia Tools and Applications*, *79*(47), 35583–35617. https://doi.org/10.1007/s11042-019-08546-5

Adekola, O. D., Maitanmi, S. O., Kasali, F. A., Omotunde, A., & Akande, O. (2020). MOVIE SUCCESS PREDICTION USING DATA MINING. *Technology, 4(2), 22-30.*. https://doi.org/10.52589/BJCNIT-CQOCIREC

Aparicio, J.T., Aparicio, M., Costa, C.J. (2023). Design Science in Information Systems and Computing. In: Anwar, S., Ullah, A., Rocha, Á., Sousa, M.J. (eds) Proceedings of International Conference on Information Technology and Applications. Lecture Notes in Networks and Systems, vol 614. Springer, Singapore. https://doi.org/10.1007/978-981-19-9331-2_35

Aparicio, J. T., de Sequeira, J. S., & Costa, C. J. (2021). Emotion analysis of portuguese political parties communication over the covid-19 pandemic. In *2021 16th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1-6). IEEE. https://doi.org/10.23919/CISTI52073.2021.9476557

Aparicio, S., Aparicio, J. T., & Costa, C. J. (2019). Data Science and AI: trends analysis. In *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1-6). IEEE. https://doi.org/10.23919/CISTI.2019.8760820

Arriaga, A & Costa, C.J. (2023). Modeling and Predicting Daily COVID-19 (SARS-CoV-2) Mortality in Portugal. In: Anwar, S., Ullah, A., Rocha, Á., Sousa, M.J. (eds) Proceedings of International Conference on Information Technology and Applications. Lecture Notes in Networks and Systems, vol 614. Springer, Singapore. https://doi.org/10.1007/978-981-19-9331-2_23

Beckh, K., Müller, S., Jakobs, M., Toborek, V., Tan, H., Fischer, R., Welke, P., Houben, S., & von Rueden, L. (2021). Explainable Machine Learning with Prior Knowledge: An Overview. Em *arXiv e-prints*. https://doi.org/10.48550/arXiv.2105.10172

Bodapati, J., Veeranjaneyulu, N., & Shaik, S. (2019). Sentiment Analysis from Movie Reviews Using LSTMs. *Ingénierie des systèmes d information*, *24*(1), 125–129. https://doi.org/10.18280/isi.240119

Bristi, W. R., Zaman, Z., & Sultana, N. (2019). Predicting IMDb Rating of Movies by Machine Learning Techniques. *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–5. https://doi.org/10.1109/ICCCNT45670.2019.8944604

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). CRISP-DM 1.0 step-by-step data mining guide. Em *Springer*.

Cheang, Y. M., & Chye Cheah, T. (2021). Predicting Movie Box-Office Success and The Main Determinants of Movie Box Office Sales in Malaysia using Machine Learning Approach. *Proceedings of the 2021 10th International Conference on Software and Computer Applications*, 57–62. https://doi.org/10.1145/3457784.3457793

Çizmeci, B., & Ögüdücü, Ş. G. (2018). Predicting IMDb Ratings of Pre-release Movies with Factorization Machines Using Social Media. *2018 3rd International Conference on Computer Science and Engineering (UBMK)*, 173–178. https://doi.org/10.1109/UBMK.2018.8566661

Costa, C. J., & Aparicio, J. T. (2020). POST-DS: A Methodology to Boost Data Science. *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)*, 1–6. https://doi.org/10.23919/CISTI49556.2020.9140932

Costa, C.J., Aparicio, J.T. (2021). A Methodology to Boost Data Science in the Context of COVID-19. In: Arabnia, H.R., et al. Advances in Parallel & Distributed Processing, and Applications. Transactions on Computational Science and Computational Intelligence. Springer, Cham. https://doi.org/10.1007/978-3-030-69984-0_7

Costa, C.J.; Aparicio, M (2023) Applications of Data Science and Artificial Intelligence. *Applied Sciences 13*, 9015. https://doi.org/10.3390/app13159015

Dietterichl, T. G. (2002). Ensemble Learning. Em M. Arbib (Ed.), *The Handbook of Brain Theory and Neural Networks* (pp. 405–408). MIT Press.

Dhir, R., & Raj, A. (2018, December). Movie success prediction using machine learning algorithms and their comparison. In 2018 first international conference on secure cyber computing and communication (ICSCCC) (pp. 385-390). IEEE.. https://doi.org/10.1109/ICSCCC.2018.8703320

Dixit, P., Hussain, S., & Singh, G. (2020). Predicting the IMDB rating by using EDA and machine learning Algorithms. Em *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*. https://doi.org/10.32628/CSEIT206481

Gomes, A. L., Vianna, G., Escovedo, T., & Kalinowski, M. (2022, maio 16). Predicting IMDb Rating of TV Series with Deep Learning: The Case of Arrow. *Anais Do Simpósio Brasileiro de Sistemas de Informação (SBSI)*. Anais do XVIII Simpósio Brasileiro de Sistemas de Informação. https://sol.sbc.org.br/index.php/sbsi/article/view/21350

Gupta, K., Bajpayee, S., & Priyadharsini, A. M. (2019). Movie Success Prediction. *International Journal of Recent Technology and Engineering (IJRTE)*, *8*(3), 5659–5663. https://doi.org/10.35940/ijrte.B2484.098319

Gupta, V., Jain, N., Garg, H., Jhunthra, S., Mohan, S., Omar, A. H., & Ahmadian, A. (2023). Predicting attributes based movie success through ensemble machine learning. *Multimedia Tools and Applications*, *82*(7), 9597–9626. https://doi.org/10.1007/s11042-021-11553-0

Hsu, P.-Y., Shen, Y.-H., & Xie, X.-A. (2014). Predicting Movies User Ratings with Imdb Attributes. Em D. Miao, W. Pedrycz, D. Ślęzak, G. Peters, Q. Hu, & R. Wang (Eds.), *Rough Sets and Knowledge Technology* (pp. 444–453). Springer International Publishing. https://doi.org/10.1007/978-3-319-11740-9_41

Hutto, C., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 216-225). https://doi.org/10.1609/icwsm.v8i1.14550

Jain, V. (2013). *Prediction of Movie Success using Sentiment Analysis of Tweets*. *3*(3).

Jaiswal, S., & Sharma, D. (2017). *Predicting Success of Bollywood Movies Using Machine Learning Techniques*. 121–124. https://doi.org/10.1145/3140107.3140126

Meenakshi, K., Maragatham, G., Agarwal, N., & Ghosh, I. (*2018, April)*. A Data mining Technique for Analyzing and Predicting the success of Movie. *In Journal of Physics: Conference Series* (Vol. 1000, No. 1, p. 012100). IOP Publishing.. https://doi.org/10.1088/1742-6596/1000/1/012100

Mhowwala, Z., Sulthana, A. R., & Shetty, S. D. (2020). Movie Rating Prediction using Ensemble Learning Algorithms. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *11*(8), Artigo 8. https://doi.org/10.14569/IJACSA.2020.0110849

Mundra, S., Dhingra, A., Kapur, A., & Joshi, D. (2019). Prediction of a Movie's Success Using Data Mining Techniques. Em S. C. Satapathy & A. Joshi (Eds.), *Information and Communication Technology for Intelligent Systems* (pp. 219–227). Springer. https://doi.org/10.1007/978-981-13-1742-2_22

Quader, N., Gani, Md. O., Chaki, D., & Ali, Md. H. (2017). A machine learning approach to predict movie box-office success. *2017 20th International Conference of Computer and Information Technology (ICCIT)*, 1–7. https://doi.org/10.1109/ICCITECHN.2017.8281839

Vr, N., Pranav, M., Babu, P., & Lijiya, A. (2014). Predicting Movie Success Based on IMDB Data. *International Journal of Business Intelligents*, *003*, 34–36. https://doi.org/10.20894/IJBI.105.003.002.004

Yoo, S., Kanter, R., Cummings, D., & Maas, A. L. (2011). *Predicting Movie Revenue from IMDb Data*. https://www.semanticscholar.org/paper/Predicting-Movie-Revenue-from-IMDb-Data-Yoo-Kanter/6e6cdf5b0282d89de45c407fc76a4c218696e3e3

Anexo

Link da página GitHub com o código:
https://github.com/javp99/filmes