MINISTÉRIO DA EDUCAÇÃO
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
DE GOIÁS - CÂMPUS GOIÂNIA

CURSO DE LICENCIATURA EM MATEMÁTICA

TÉCNICAS DE *MACHINE LEARNING* PARA ESTIMAÇÃO DE FATURAMENTO COM O JOGO LEAGUE OF LEGENDS

SILAS BATISTA DE URZEDA

GOIÂNIA - GOIÁS
2024

TERMO DE AUTORIZAÇÃO PARA DISPONIBILIZAÇÃO NO REPOSITÓRIO DIGITAL DO IFG - ReDi IFG

Com base no disposto na Lei Federal nº 9.610/98, AUTORIZO o Instituto Federal de Educação, Ciência e Tecnologia de Goiás, a disponibilizar gratuitamente o documento no Repositório Digital (ReDi IFG), sem ressarcimento de direitos autorais, conforme permissão assinada abaixo, em formato digital para fins de leitura, download e impressão, a título de divulgação da produção técnico-científica no IFG.

Identificação da Produção Técnico-Científica

[ ] Tese	[ ] Artigo Científico
[ ] Dissertação	[ ] Capítulo de Livro
[ ] Monografia – Especialização	[ ] Livro
[X] TCC - Graduação	[ ] Trabalho Apresentado em Evento
[ ] Produto Técnico e Educacional - Tipo: _____	

Nome Completo do Autor: Silas Batista Urzêda
Matrícula: 20211010940149
Título do Trabalho: TÉCNICAS DE MACHINE LEARNING PARA ESTIMAÇÃO DE FATURAMENTO COM O JOGO LEAGUE OF LEGENDS

Autorização - Marque uma das opções

1. (X) Autorizo disponibilizar meu trabalho no Repositório Digital do IFG (acesso aberto);
2. ( ) Autorizo disponibilizar meu trabalho no Repositório Digital do IFG somente após a data ___/___/_____ (Embargo);
3. ( ) Não autorizo disponibilizar meu trabalho no Repositório Digital do IFG (acesso restrito).

Ao indicar a opção **2 ou 3**, marque a justificativa:

( ) O documento está sujeito a registro de patente.
( ) O documento pode vir a ser publicado como livro, capítulo de livro ou artigo.
( ) Outra justificativa: _____

DECLARAÇÃO DE DISTRIBUIÇÃO NÃO-EXCLUSIVA

O/A referido/a autor/a declara que:

i. o documento é seu trabalho original, detém os direitos autorais da produção técnico-científica e não infringe os direitos de qualquer outra pessoa ou entidade;
ii. obteve autorização de quaisquer materiais inclusos no documento do qual não detém os direitos de autor/a, para conceder ao Instituto Federal de Educação, Ciência e Tecnologia de Goiás os direitos requeridos e que este material cujos direitos autorais são de terceiros, estão claramente identificados e reconhecidos no texto ou conteúdo do documento entregue;
iii. cumpriu quaisquer obrigações exigidas por contrato ou acordo, caso o documento entregue seja baseado em trabalho financiado ou apoiado por outra instituição que não o Instituto Federal de Educação, Ciência e Tecnologia de Goiás.

Goiânia, 30/09/2024.
Local Data

Silas Batista de Urzeda
Assinatura do Autor e/ou Detentor dos Direitos Autorais

SILAS BATISTA DE URZEDA

TÉCNICAS DE *MACHINE LEARNING* PARA ESTIMAÇÃO DE FATURAMENTO COM O JOGO LEAGUE OF LEGENDS

Trabalho de Conclusão de Curso apresentado ao Curso de Licenciatura em Matemática do Instituto Federal de Goiás - IFG - Câmpus Goiânia, como requisito parcial para obtenção do título de Licenciado em Matemática.

Área de Concentração: Matemática

Orientador: Prof. Doutor Uender Barbosa de Souza

GOIÂNIA - GOIÁS
2024

U83t Urzeda, Silas Batista de.
Técnicas de Machine Learning para estimação de faturamento com o jogo League of Legends / Silas Batista de Urzeda. – Goiânia: Instituto Federal de Educação, Ciência e Tecnologia de Goiás, Câmpus Goiânia, 2024.
71 f.: il.

Orientador: Prof. Dr. Uender Barbosa de Souza.

TCC (Trabalho de Conclusão de Curso) – Licenciatura em Matemática, Instituto Federal de Educação, Ciência e Tecnologia de Goiás, Câmpus Goiânia.

1. Machine Learning. 2. League of Legends (Jogo) - faturamento. 3. Estatística - análise de dados. I. Souza, Uender Barbosa de (orientador). II. Instituto Federal de Educação, Ciência e Tecnologia de Goiás, Câmpus Goiânia. III. Título.

CDD 519.5

Ficha catalográfica elaborada pelo Bibliotecário Alisson de Sousa Belthodo Santos CRB1/ 2.266
Biblioteca Professor Jorge Félix de Souza,
Instituto Federal de Educação, Ciência e Tecnologia de Goiás, Câmpus Goiânia.

MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE GOIÁS
CÂMPUS GOIÂNIA

TÉCNICAS DE *MACHINE LEARNING* PARA ESTIMAÇÃO DE FATURAMENTO COM O JOGO *LEAGUE OF LEGENDS*

por

SILAS BATISTA DE URZEDA

Trabalho de Conclusão de Curso apresentado ao Colegiado do Curso de Licenciatura em Matemática do Instituto Federal de Educação, Ciência e Tecnologia de Goiás – Câmpus Goiânia, como requisito parcial para a obtenção do Diploma de Licenciado em Matemática.

Área de concentração: Matemática

Orientador: Prof. Dr. Uender Barbosa de Souza

Trabalho de Conclusão de Curso defendido e aprovado, em 25 de setembro de 2024, pela banca examinadora constituída pelos seguintes membros:

(assinado eletronicamente)

Prof. Dr. Uender Barbosa de Souza(Presidente da Banca/IFG/Câmpus Goiânia)

***(assinado* eletronicamente)**

Prof. Dr. Jose Elmo de Menezes (IFG/Câmpus Goiânia)

(assinado eletronicamente)

Prof. Me. Ricardo da Silva Santos (IFG/Câmpus Goiânia)

Documento assinado eletronicamente por:

- **Jose Elmo de Menezes**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 25/09/2024 18:00:02.
- **Ricardo da Silva Santos**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 25/09/2024 17:59:50.
- **Uender Barbosa de Souza**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 25/09/2024 17:58:44.

Este documento foi emitido pelo SUAP em 10/09/2024. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifg.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código Verificador: 562295
Código de Autenticação: 975578df7b

Agradecimentos

A realização deste trabalho só foi possível graças ao apoio e à colaboração de diversas pessoas, às quais devo expressar meu mais profundo agradecimento.

No âmbito acadêmico, sou imensamente grato ao meu orientador Prof. Dr. Uender Barbosa de Souza, que com extrema paciência e dedicação me guiou na elaboração e adaptação dos conceitos específicos e aplicados deste estudo. Sua coragem e disposição em enfrentar um tema tão inédito e moderno quanto o meu foram fundamentais para o desenvolvimento deste trabalho.

Expresso também meus sinceros agradecimentos aos membros da banca avaliadora. Ao Prof. Dr. José Elmo de Menezes, cujas aulas de Estatística despertaram em mim o pensamento crítico e a habilidade de identificar padrões em diferentes conjuntos de dados, além de me introduzir ao uso da ferramenta R, essencial para a programação de operações complexas relacionadas à análise de dados. Ao Prof. Dr. Márcio Dias de Lima, cuja Prática como Componente Curricular (PCC) em *Python* me revelou a simplicidade e versatilidade desta linguagem, abrindo um leque de possibilidades. E ao Prof. Me. Ricardo da Silva Santos, que, conduziu com excelência as aulas remotas de Álgebra Linear durante o período pandêmico.

Gostaria de expressar minha gratidão também aos colegas de curso, com quem, infelizmente, tive pouco contato devido à carga intensa de até oito disciplinas em diversos períodos. Ainda assim, aqueles com quem pude compartilhar momentos foram sempre muito cordiais e colaborativos. Agradeço, igualmente, aos amigos das turmas de Licenciatura em Letras, História e Física, que me acompanharam nos mais variados momentos e que estão presentes não apenas na minha trajetória acadêmica, mas também na minha vida pessoal.

Por fim, mas não menos importante, meus agradecimentos à tríade essencial: Deus, minha família e minha psiquiatra. A Deus, por me conceder a graça da vida e a oportunidade de alcançar meus sonhos; à minha família, por todo o apoio contínuo, tanto em minha vida pessoal quanto acadêmica, especialmente nos momentos mais desafiadores; e à minha psiquiatra, que me acompanha há mais de cinco anos, proporcionando-me o su-

porte necessário para superar as adversidades e retomar meus estudos, permitindo que eu conclua este projeto que sempre almejei.

Resumo

Este trabalho visa desenvolver um modelo preditivo capaz de estimar o faturamento anual da Riot Games com o jogo League of Legends (LoL), utilizando técnicas de *Machine Learning* (ML) e dados públicos relacionados aos conteúdos lançados no jogo. LoL é um dos jogos eletrônicos mais populares do mundo, operando em um modelo *free-to-play* (grátis para jogar) que monetiza principalmente por meio da venda de itens cosméticos, como *skins*. Apesar do sucesso comercial, a Riot Games não divulga detalhadamente seus dados financeiros, tornando desafiadora a compreensão do impacto econômico de suas estratégias de conteúdo. Para atingir o objetivo proposto, foram coletados e organizados dados históricos sobre campeões, *skins* e eventos a partir de fontes públicas confiáveis, como o *League of Legends Fandom*. O pré-processamento dos dados incluiu a transformação em vetores de características adequados para aplicação em modelos de aprendizado de máquina. Foram implementados e comparados modelos de Regressão Linear, Árvore de Decisão e *Random Forest*, avaliando seu desempenho na previsão do faturamento. Uma técnica de redução de dimensionalidade, a Análise de Componentes Principais (PCA), foi aplicada para otimizar os modelos e melhorar a interpretabilidade dos resultados. Os resultados indicaram que o modelo *Random Forest*, especialmente quando combinado com técnicas de normalização e PCA, apresentou o melhor desempenho na previsão do faturamento. Apesar das limitações decorrentes da utilização de estimativas de faturamento obtidas em sites de notícias e fóruns, os modelos conseguiram captar tendências gerais, oferecendo *insights* valiosos sobre os fatores que influenciam a receita da Riot Games. O estudo evidencia a viabilidade de aplicar técnicas de *Machine Learning* na previsão de faturamento em contextos com dados limitados ou parcialmente indisponíveis.

Palavras-chave: *Machine Learning*; Previsão de Faturamento; League of Legends; *Random Forest*; Redução de Dimensionalidade.

Abstract

This work aims to develop a predictive model capable of estimating Riot Games' annual revenue from the game League of Legends (LoL), using Machine Learning (ML) techniques and public data related to the content released in the game. LoL is one of the most popular electronic games in the world, operating on a free-to-play model that monetizes mainly through the sale of cosmetic items such as skins. Despite its commercial success, Riot Games does not disclose its financial data in detail, making it challenging to understand the economic impact of its content strategies. To achieve the proposed objective, historical data on champions, skins, and events was collected and organized from reliable public sources, such as League of Legends Fandom. Data pre-processing included transformation into feature vectors suitable for application in machine learning models. Linear Regression, Decision Tree and Random Forest models were implemented and compared, assessing their performance in predicting revenue. A dimensionality reduction technique, Principal Component Analysis (PCA), was applied to optimize the models and improve the interpretability of the results. The results indicated that the Random Forest model, especially when combined with normalization and PCA techniques, showed the best performance in predicting turnover. Despite the limitations arising from the use of revenue estimates obtained from news sites and forums, the models managed to capture general trends, offering valuable insights into the factors that influence Riot Games' revenue. The study demonstrates the feasibility of applying Machine Learning techniques to revenue forecasting in contexts with limited or partially unavailable data.

Keywords: Machine Learning; Revenue Prediction; League of Legends; Random Forest; Dimensionality Reduction.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

No cenário contemporâneo dos jogos eletrônicos, poucos títulos alcançaram o impacto e a popularidade de League of Legends (LoL). Desde seu lançamento em 2009 pela Riot Games, o jogo estabeleceu-se como um fenômeno global, redefinindo o panorama dos esportes eletrônicos e influenciando a cultura *gamer* ao redor do mundo. Operando em um modelo de negócios *free-to-play*, League of Legends permite acesso gratuito ao jogo, monetizando-se principalmente por meio da venda de itens cosméticos, conhecidos como *skins*. Essa estratégia tem se mostrado altamente eficaz, permitindo à Riot Games alcançar receitas significativas sem cobrar diretamente pelo acesso ao jogo.

Entretanto, apesar da visibilidade e do sucesso comercial, a Riot Games não divulga detalhadamente seus dados financeiros, tornando desafiadora a compreensão do impacto econômico de suas estratégias de conteúdo. A falta de transparência em relação ao faturamento específico, especialmente associado a elementos como lançamentos de novos campeões, *skins* e eventos sazonais, dificulta análises aprofundadas sobre como esses fatores influenciam a receita da empresa.

Diante desse contexto, para compreender as estratégias da empresa surge a necessidade de desenvolver métodos que possibilitem estimar o faturamento anual utilizando dados públicos disponíveis. Vale ressaltar que a previsão de receitas é uma prática essencial em diversos setores, auxiliando no planejamento estratégico, na alocação de recursos e na tomada de decisões informadas. No setor de jogos eletrônicos, essa compreensão é ainda mais crucial, dado o ritmo acelerado de inovação e a intensa competitividade.

O problema central deste trabalho consiste, portanto, em prever o faturamento anual da Riot Games com League of Legends a partir de variáveis públicas, como o número de campeões e *skins* lançados, características dos campeões, eventos e outras métricas relacionadas ao jogo. Essa tarefa apresenta desafios significativos devido à ausência de dados financeiros oficiais e à complexidade inerente ao mercado de jogos eletrônicos.

Justifica-se este estudo pela relevância econômica e cultural de League of Legends, bem como pela contribuição acadêmica que a aplicação de técnicas de Aprendizado de Máquina (*Machine Learning*) pode oferecer à análise de dados nesse contexto. Estimar o faturamento com base em variáveis acessíveis não apenas preenche a lacuna de informação financeira para compreensão das estratégias de mercado, mas também demonstra a viabilidade de modelos preditivos em cenários com dados limitados ou parcialmente indisponíveis.

O objetivo geral deste trabalho é desenvolver um modelo preditivo capaz de estimar o faturamento anual da Riot Games com League of Legends, utilizando técnicas de *Machine Learning* e dados históricos relacionados aos conteúdos lançados no jogo. Para atingir esse

propósito, foram estabelecidos os seguintes objetivos específicos:

- Coletar e organizar dados históricos sobre campeões, *skins* e eventos de League of Legends a partir de fontes públicas confiáveis.
- Pré-processar os dados, transformando-os em vetores de características adequados para a aplicação de modelos de aprendizado de máquina.
- Implementar e comparar diferentes modelos preditivos, como Regressão Linear, Árvore de Decisão e *Random Forest*, avaliando seu desempenho na previsão do faturamento.
- Aplicar técnicas de redução de dimensionalidade, como a Análise de Componentes Principais (PCA), para otimizar os modelos e melhorar a interpretabilidade dos resultados.
- Analisar os resultados obtidos, identificando os fatores que mais influenciam o faturamento e avaliando a eficácia dos modelos propostos.

A metodologia adotada combina técnicas quantitativas de análise de dados com a aplicação de algoritmos de *Machine Learning*. Os dados foram coletados de fontes públicas, como o League of Legends Fandom, que fornece informações detalhadas sobre lançamentos de campeões, *skins* e eventos. As estimativas de faturamento foram obtidas de sites especializados e fóruns da comunidade, considerando a indisponibilidade de dados oficiais.

Este estudo delimita-se ao período entre 2015 e 2024, devido à indisponibilidade de estimativas de faturamento anteriores e à relevância dos lançamentos nesse intervalo temporal. A análise concentra-se exclusivamente em League of Legends, não abrangendo outros títulos ou iniciativas da Riot Games.

O trabalho está estruturado em cinco capítulos. Primeiro, esta introdução. Em seguida, o referencial teórico aborda conceitos fundamentais sobre League of Legends, modelos de negócios em jogos eletrônicos, fundamentos de *Machine Learning* e técnicas estatísticas relevantes. A metodologia detalha o processo de coleta e pré-processamento dos dados, os modelos implementados e as técnicas de validação utilizadas. No capítulo de desenvolvimento e resultados, são apresentados os experimentos realizados, acompanhados de análises críticas. Por fim, as conclusões destacam as contribuições do estudo, suas limitações e sugestões para trabalhos futuros.

Espera-se que este trabalho contribua para pesquisadores que busquem uma introdução ao conhecimento na área de previsão de faturamento em jogos eletrônicos, demonstrando a aplicabilidade de técnicas de *Machine Learning* em contextos com dados limitados. Além disso, busca-se oferecer *insights* úteis para profissionais da indústria, acadêmicos e entusiastas, ampliando a compreensão sobre os fatores que influenciam o sucesso financeiro de jogos como League of Legends.

2 Referencial Teórico

Este capítulo destina-se à exposição dos conceitos e definições fundamentais que embasam o desenvolvimento do presente estudo. Serão abordados aspectos como a estrutura e funcionamento do jogo objeto da pesquisa, e seu modelo de negócios, bem como uma breve introdução à Inteligência Artificial (IA) e ao *Machine Learning* (ML), destacando suas aplicações. Para detalhes da teoria do método utilizado, o leitor pode consultar James (2013), Haslwanter (2022), Mohri (2018) e Breiman (2001). É importante deixar claro que o foco deste estudo foi a implementação dos modelos usando as bibliotecas *Python*. Além disso, foram discutidos trabalhos acadêmicos que, embora não abranjam a totalidade do tema proposto, oferecem contribuições relevantes tanto no contexto do ML quanto em estudos relacionados ao jogo.

2.1 O Jogo

League of Legends (LoL), um dos jogos mais influentes no gênero *Multiplayer Online Battle Arena* (MOBA), surgiu como fruto de uma visão empreendedora de Marc Merrill e Brandon Beck. Ambos formados pela Universidade do Sul da Califórnia (USC), os dois eram ávidos entusiastas de jogos da Blizzard Entertainment, especialmente Warcraft III e Starcraft. Merrill e Beck perceberam uma oportunidade ao observar o crescente sucesso de modificações (*mods*) como Defense of the Ancients (DotA), um *mod* popular de Warcraft III. O processo de jogar DotA, no entanto, exigia que os jogadores adquirissem o jogo base (Warcraft III), o que se configurava como uma barreira de entrada para novos usuários. Identificando essa limitação, Merrill e Beck idealizaram um jogo que oferecesse uma experiência similar à de DotA, mas com acesso gratuito, acreditando que isso poderia atrair um público muito maior.

A Riot Games, fundada em 2006 com um investimento inicial de US$ 1,5 milhão, foi a plataforma por meio da qual Merrill e Beck materializaram essa visão. A empresa reuniu desenvolvedores com experiência em *mods*, o que possibilitou um desenvolvimento técnico de alta qualidade. Após anos de desenvolvimento, League of Legends foi lançado oficialmente em 2009. Apesar de um lançamento inicial modesto nos Estados Unidos, a decisão estratégica de firmar uma parceria com a gigante chinesa Tencent para a distribuição na China foi um divisor de águas para o crescimento do jogo. Em questão de meses, League of Legends acumulou mais de 100 mil jogadores na China, fenômeno que contribuiu para a expansão de sua popularidade global (Corrêa, 2020).

A escolha do modelo de negócio *free-to-play* (grátis para jogar) com monetização ba-

seada em itens cosméticos, como *skins* e outros elementos visuais, foi uma inovação que se provou bem-sucedida e revolucionária na indústria de jogos online. Esse modelo permitiu que milhões de jogadores acessassem o jogo sem qualquer custo inicial, enquanto gerava receitas substanciais por meio de compras *in-game* (dentro do jogo), transformando a Riot Games em uma das maiores desenvolvedoras do mundo. A comunidade de jogadores, além de ser o público-alvo, desempenhou um papel essencial no crescimento contínuo e na adaptação do jogo a novos mercados, provando a importância da interação entre desenvolvedores e usuários para o sucesso a longo prazo de jogos *multiplayer*.

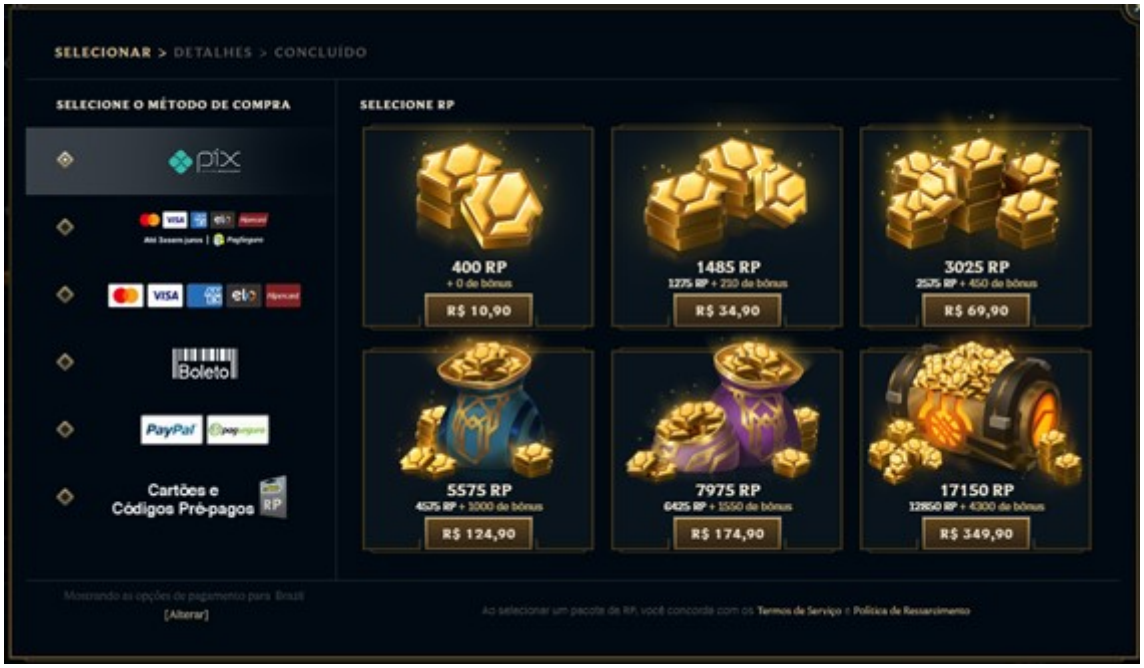
2.1.1 Modelo de Negócio

O modelo de receita do jogo League of Legends (LoL) é fundamentado na comercialização de cosméticos, conhecidos como *skins*. Cada partida envolve dois times de cinco jogadores, posicionados estrategicamente, cujo objetivo é destruir a base adversária. Os jogadores escolhem entre mais de 160 personagens disponíveis, denominados campeões. A obtenção desses campeões ocorre principalmente por meio da moeda virtual do jogo, chamada Essência Azul (EA), que pode ser adquirida gratuitamente à medida que os jogadores sobem de nível. Cada campeão possui uma aparência padrão que reflete sua história no universo do jogo.

No entanto, os jogadores têm a opção de personalizar os campeões adquirindo *skins*, que alteram a aparência visual do personagem. Essas skins são compradas com a moeda paga do jogo, os Riot Points (RP). Há uma ampla variedade de *skins*, organizadas por níveis e preços, sendo que à medida que o preço aumenta, mais detalhes visuais e efeitos personalizados são adicionados.

A obtenção de RP é realizada predominantemente por meio do *Client* do jogo (*software* onde os jogadores podem buscar partidas e realizar compras). A Figura 2.1 mostra a loja virtual onde os RP podem ser adquiridos:

Figura 2.1: Loja dentro do *Client* do jogo



Os preços diferem a medida da quantidade comprada de RP. Pode parecer pouco, mas a diferença entre o pacote mais barato e o mais caro é de 25%, como mostra a Tabela 2.1:

Tabela 2.1: Valores dos pacotes de RP. Consulta dia 02/07/2024

Quantidade de RP	Valor	Valor por RP	Economia em Relação ao Pacote Mais Barato
400	R$ 10,90	R$ 0,0273	-
1485	R$ 34,90	R$ 0,0235	13,8%
3025	R$ 69,90	R$ 0,0231	15,2%
5575	R$ 124,90	R$ 0,0224	17,8%
7975	R$ 174,90	R$ 0,0219	19,5%
17150	R$ 349,90	R$ 0,0204	25,1%

As skins são classificadas em seis *tiers* (categorias): Comuns, Épicas, Lendárias, Míticas, *Ultimate*, Variante Mítica e Transcendente. A Tabela 2.2 a seguir apresenta o valor máximo de cada *tier*, com base no valor médio por RP de R$ 0,0231:

Tabela 2.2: Valor médio em reais de cada tipo de *skin*

Tier	Valor em RP	Valor
Comum	975	R$ 22,52
Épica	1350	R$ 31,18
Lendária	1820	R$ 42,04
Mítica	2200	R$ 50,82
Ultmate	3250	R$ 75,07
Variante Mítica	22500	R$ 519,74
Transcendente	59260	R$ 1.368,87

Abaixo, estão as diferenças visuais entre o modelo base da campeã Miss Fortune e suas *skins* de nível superior:

Figura 2.2: Miss Fortune modelo base

Figura 2.3: Miss Fortune (Mítica)

Figura 2.4: Miss Fortune (Variante Mítica)

Em estudo realizado em 2014, o analista da Ubisoft, Teut Weidemann, contou a do site Game Developer (Alexander, 2023) que observou uma discrepância entre a taxa de monetização de League of Legends e a média da indústria de *games*. Enquanto apenas cerca de 4% dos jogadores adquiriam itens cosméticos, um percentual consideravelmente inferior à faixa de 15% a 25% observada no mercado como um todo, Weidemann argumentou que a lucratividade do jogo era sustentada principalmente por sua vasta base de jogadores.

Dados subsequentes corroboram essa análise: em 2017, League of Legends gerou uma receita de US$ 2,1 bilhões (Gill, 2023); em 2018, embora tenha apresentado uma leve retração para US$ 1,4 bilhões, o título manteve sua posição como um dos jogos mais lucrativos do mercado (SuperData, 2018). Nos anos seguintes, a receita apresentou recuperação, atingindo US$ 1,5 bilhão em 2019 e US$ 1,75 bilhão em 2020 (Gill, 2023).

A longevidade do jogo é evidenciada pela alta taxa de engajamento dos jogadores, com mais de três bilhões de horas jogadas por mês em 2016, conforme reportado pela revista REUTERS (MEDIAs, 2021). Esses dados demonstram a capacidade de League of Legends em gerar receita significativa a partir de uma base de jogadores massiva, mesmo com uma taxa de monetização por jogador relativamente baixa quando comparada à média do mercado.

2.2 Estudos já Realizados sobre o Tema

A aplicação de técnicas de *Machine Learning* em diversas áreas tem sido objeto de estudo em inúmeros trabalhos acadêmicos. No contexto de previsão de séries temporais e análise de dados financeiros, bem como na indústria de jogos eletrônicos, diversos autores têm explorado o potencial dessas técnicas para melhorar a precisão de predições e auxiliar na tomada de decisões.

Em *Estudo Comparativo entre Algoritmos de Machine Learning Aplicados à Previsão de Séries Temporais do Mercado Financeiro*, os pesquisadores Alves e Prado (2022), investigam a aplicação de três modelos de *Machine Learning* na predição de dados futuros relacionados a séries temporais do mercado financeiro. Entre os algoritmos analisados, o modelo de regressão linear apresentou um desempenho satisfatório. Em outro estudo, intitulado *Técnicas de Machine Learning Aplicadas na Recuperação de Crédito do Mercado Brasileiro* (Forti, 2018), os autores identificam que os modelos *Gradient Boosting* e *Support Vector Machine* proporcionaram melhores resultados em comparação com a regressão logística, que, segundo os autores, é amplamente utilizada no setor financeiro.

No que se refere à aplicação de *Machine Learning* na análise de rentabilidade em jogos, a literatura existente é ainda limitada. No entanto, alguns estudos fazem a conexão entre *Machine Learning* e a previsão de resultados em partidas. Arık (2023), por exemplo, em *Machine Learning Models on MOBA Gaming: League of Legends Winner Prediction*, explora o uso de modelos de ML para prever o time vencedor em partidas de LoL. Já Smit (2019), em *A Machine Learning Approach for Recommending Items in League of Legends*, discute a aplicação de ML na recomendação de itens comprados por jogadores durante uma partida, correlacionando essas escolhas à probabilidade de vitória.

Em suma, embora existam estudos que abordam a aplicação de *Machine Learning* em contextos financeiros e na previsão de resultados em jogos eletrônicos, observa-se uma lacuna na literatura quanto à utilização dessas técnicas para prever o faturamento de jogos *free-to-play*. Posto isto, justifica-se este trabalho visando contribuir com o preenchimento dessa lacuna.

2.3 Principais Conceitos

Nesta seção serão introduzidos os conceitos e métodos fundamentais para o desenvolvimento deste trabalho.

2.3.1 Inteligência Artificial

A previsão de escrita na barra de pesquisa do Google, tradutores automáticos, sugestões de publicações em *feeds* do Instagram e TikTok, recomendações de vídeos no YouTube e Netflix, carros autônomos e o uso de sistemas como o ChatGPT são todos elementos representativos de uma realidade contemporânea moldada pela Inteligência Artificial (IA).

Na obra de Russell e Norvig (2016), a IA é apresentada como o estudo de técnicas voltadas para capacitar os computadores a agirem racionalmente. Para os autores, agir racionalmente significa maximizar a utilidade esperada, isto é, tomar decisões que aumentem a probabilidade de alcançar um objetivo previamente estabelecido. A IA, conforme proposto, abrange um conjunto amplo de capacidades cognitivas, incluindo a aprendizagem, o raciocínio lógico, o planejamento de ações, a percepção do ambiente, o processamento de linguagem natural e a interação física, principalmente em robótica.

Russell e Norvig (2016) também sugerem que a IA pode ser categorizada em quatro perspectivas principais. A primeira delas trata dos sistemas que pensam como humanos, ou seja, que buscam simular os processos cognitivos humanos, como a aprendizagem e o raciocínio. A segunda perspectiva foca em sistemas que agem como humanos, onde o objetivo é desenvolver máquinas capazes de exibir comportamentos semelhantes aos dos seres humanos, sendo o Teste de Turing uma referência importante para avaliar essa capacidade. Já a terceira perspectiva aborda sistemas que pensam racionalmente, fundamentados em lógica matemática e raciocínio formal para otimizar a tomada de decisões. Por fim, a quarta perspectiva, que trata dos sistemas que agem racionalmente, enfatiza a maximização das chances de sucesso em tarefas específicas, alinhando as ações das máquinas aos objetivos concretos.

2.3.2 *Machine Learning*

No campo do *Machine Learning* (ML) ou Aprendizado de Máquina, conforme definido por Mohri (2018), trata-se do estudo de algoritmos e modelos estatísticos que permitem que sistemas computacionais realizem tarefas sem depender de instruções programadas para cada situação. Samuel (1959), pioneiro no campo, descreve ML como a capacidade de os computadores aprenderem sem serem explicitamente programados para isso. Em uma visão mais recente, Brown (2021) em um artigo para Instituto de Tecnologia de Massachusetts (MIT) define ML como um subcampo da IA, capaz de criar máquinas que imitam

a inteligência humana ao resolver problemas de maneira semelhante à nossa. O sucesso do ML depende essencialmente de dados reitera Brown, como números de vendas, fotos de pessoas em *shoppings* ou dados de audiência em programas de TV. Quanto mais dados são fornecidos, mais eficaz se torna a análise realizada pelo sistema. Posteriormente, que o programa deve escolher o modelo de aprendizado a ser utilizado.

> A função de um sistema de aprendizado de máquina pode ser descritiva, o que significa que o sistema usa os dados para explicar o que aconteceu; preditiva, o que significa que o sistema usa os dados para prever o que vai acontecer; ou prescritiva, o que significa que o sistema usará os dados para fazer sugestões sobre quais ações toma (Malone et al., 2020, p. 16, tradução própria).

Três modelos podem ser definidos entre as tais funções mencionadas:

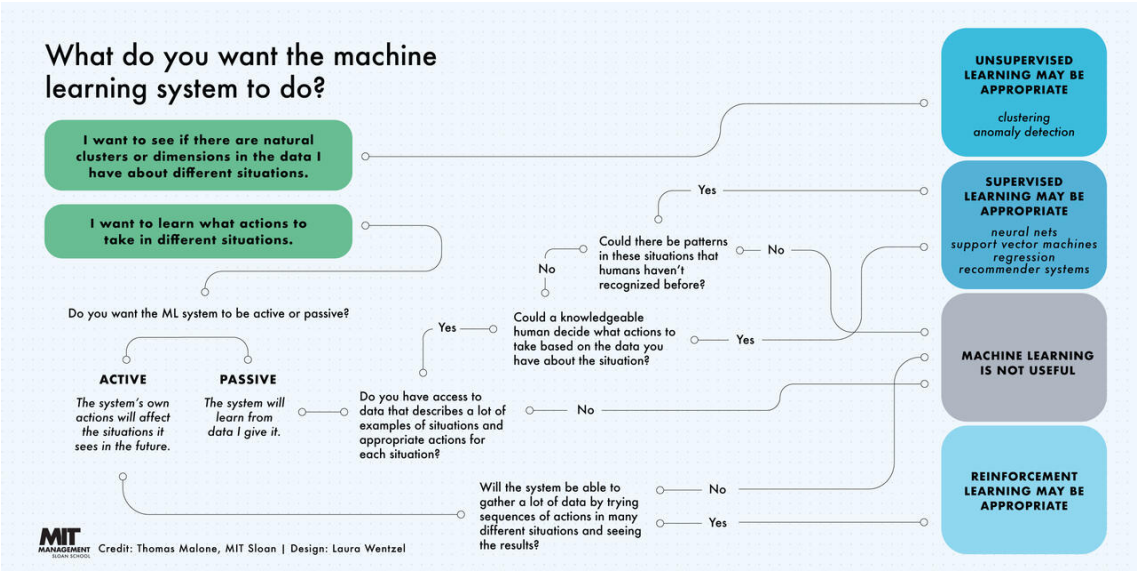
(i) Modelos supervisionados: são treinados com conjuntos de dados rotulados, como as relações entre a quantidade de chuva e a umidade de uma região.

(ii) Modelos não supervisionados: analisam dados não rotulados, buscando padrões ocultos, como o comportamento de jogadores que compram itens em jogos online.

(iii) Modelos de reforço: aprendem por tentativa e erro, visando maximizar recompensas e minimizar punições, sendo amplamente usados em carros autônomos e robôs jogadores de xadrez.

O fluxograma abaixo resume a utilização e melhor modelo que pode ser usado para cada objetivo:

Figura 2.5: Fluxograma sobre utilização de ML



Fonte: Brown (2021).

Além desses, o ML é aplicado em várias outras áreas da IA, como no processamento de linguagem natural (utilizado por assistentes virtuais como Google Assistant e Siri), redes neurais (capazes de classificar sons, como o canto de uma cigarra), e na aprendizagem profunda (*deep learning*), usada em carros autônomos e diagnósticos médicos.

Uma pesquisa de Brynjolfsson et al. (2018) sugere que a IA e o ML terão impacto em todas as áreas da sociedade, mas sem substituir completamente os profissionais. Em vez disso, essas tecnologias tendem a atuar em funções que exigem excesso de trabalho humano, como a análise de bilhões de imagens para identificar indivíduos procurados pela justiça.

2.3.3 Previsão e Projeção

Historicamente, as pessoas buscavam o Oráculo em Delfos, na Grécia, para obter previsões sobre seu futuro, enquanto meteorologistas analisavam órgãos de animais, como fígados de ovelhas, na tentativa de prever eventos futuros. Hyndman (2018) explora o interesse pelas previsões desde a antiguidade. Conforme o autore, a previsão é definida como o esforço de antecipar o futuro com o máximo de exatidão, utilizando todas as informações disponíveis, incluindo dados históricos e outros conhecimentos relevantes que impactem os resultados.

Já as projeções conforme o Public Company Accounting Oversight Board (PCAOB) (2001), são estimativas que, conforme o conhecimento e crença da parte responsável,

baseiam-se em uma ou mais suposições hipotéticas e delineiam a posição financeira, resultados operacionais, entre outros. Tais projeções são frequentemente elaboradas para explorar cenários hipotéticos, respondendo a questões como "O que aconteceria se...?". Projeções financeiras, portanto, refletem as expectativas da parte responsável quanto às condições futuras e os cursos de ação esperados, com base em suposições específicas.

A diferença entre previsão e projeção está no fato de que a primeira se baseia em dados passados, enquanto a segunda considera possíveis cenários futuros. Não se pode afirmar que uma abordagem seja mais precisa ou científica que a outra, pois ambas oferecem perspectivas complementares essenciais para um entendimento mais completo do fenômeno em análise (receita de uma empresa de jogos?).

2.3.4 Regressão Linear

A regressão linear constitui uma técnica estatística amplamente utilizada para modelar a relação entre uma variável dependente (ou resposta) e uma ou mais variáveis independentes (também denominadas preditoras). O propósito central desta metodologia é determinar a reta que melhor descreve essa relação, viabilizando a previsão de valores da variável dependente a partir dos valores observados das variáveis independentes. A seguir, serão discutidos em detalhe os principais conceitos envolvidos nesse processo.

Regressão Linear Simples

De acordo com James (2013), a regressão linear simples é uma técnica direta para prever uma resposta quantitativa Y com base em uma única variável preditora X. Essa abordagem assume haver uma relação aproximadamente linear entre X e Y. A relação pode ser expressa matematicamente pela equação:

$$Y = \beta_0 + \beta_1 X \tag{2.1}$$

Na Equação 2.1, β_0 e β_1 são constantes desconhecidas que representam, respectivamente, a interceptação e a inclinação do modelo linear. Esses valores, em conjunto, são chamados de coeficientes ou parâmetros do modelo. Com base nos dados de treinamento, é possível estimar β_0 e β_1 para realizar previsões, como, por exemplo, o peso de uma pessoa em relação a sua altura. A equação da previsão pode ser escrita como:

$$\hat{y} = \beta_0 + \beta_1 x \tag{2.2}$$

Aqui, $\hat{y}$ representa a previsão de Y quando $X = x$, e o símbolo $\hat{}$ é usado para denotar

o valor estimado de um parâmetro desconhecido.

Estimando os Coeficientes da Regressão Linear Simples

Na prática, os coeficientes β_0 e β_1 são desconhecidos. Assim, antes de utilizar a Equação 2.1 para realizar previsões, é necessário estimar esses coeficientes com base nos dados observados. Considere os pares de observações $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, onde cada par corresponde a uma medição de X e Y. O objetivo é encontrar as estimativas dos coeficientes $\hat{\beta}_0$ e $\hat{\beta}_1$ de modo que o modelo linear expresso pela Equação 2.1 ajuste-se da melhor forma possível aos dados. A equação do modelo linear ajustado é dada por:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{2.3}$$

Regressão Linear Múltipla

Conforme James (2013), ao invés de ajustar um modelo de regressão linear simples para cada variável preditora, uma abordagem mais eficaz é expandir o modelo da Equação 2.1 para acomodar diretamente múltiplos fatores preditivos. Isso pode ser feito atribuindo a cada preditor um coeficiente de inclinação distinto em um único modelo. De maneira geral, suponhamos que haja p preditores distintos. O modelo de regressão linear múltipla assume a seguinte forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{2.4}$$

Estimando os Coeficientes da Regressão Linear Múltipla

Assim como na regressão linear simples, na prática, os coeficientes $\beta_0, \beta_1, \ldots, \beta_p$ no modelo da Equação 2.4 são desconhecidos e devem ser estimados a partir dos dados. Uma vez obtidas as estimativas $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, é possível realizar previsões utilizando a fórmula:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_p X_p$$

2.3.5 Árvore de Decisão

As Árvores de Decisão constituem uma metodologia amplamente empregada para a classificação de dados no âmbito da Mineração de Dados. Essas estruturas podem ser integradas à tecnologia de indução de regras, diferenciando-se, contudo, por serem as únicas a organizar os resultados de maneira hierárquica, com um claro processo de priorização. Nesse modelo, o atributo de maior relevância é posicionado no primeiro nó da árvore,

enquanto os atributos com menor impacto são alocados nos nós subsequentes. Uma das principais vantagens desse método reside na sua capacidade de facilitar o processo decisório, destacando os fatores mais pertinentes para a análise. Ademais, as Árvores de Decisão apresentam um formato intuitivo, sendo acessíveis e compreensíveis para um público amplo. Ao ordenar os atributos por importância, essas árvores permitem que os usuários identifiquem os elementos que exercem maior influência em suas atividades, auxiliando na focalização dos fatores críticos para a tomada de decisões.

Conforme observado por Garcia (2003), as Árvores de Decisão são compostas por nós, que indicam os atributos a serem analisados, e arcos ou ramos, que emergem desses nós e representam os possíveis valores desses atributos. Cada ramo descendente reflete um valor possível atribuído a um determinado atributo. Além disso, as Árvores de Decisão incluem nós folha, os quais correspondem às diferentes classes presentes em um conjunto de treinamento. Dessa forma, cada folha está diretamente associada a uma classe específica. Ademais, cada trajeto delineado na árvore, que se estende desde a raiz até uma folha, define uma regra de classificação, evidenciando as relações entre atributos e suas respectivas classes de saída.

A Árvore de Decisão é um recurso amplamente empregado para modelar o processo de tomada de decisão, representando o tempo necessário para que o tomador de decisões chegue a uma conclusão. Conforme descrito por Gomes e Gomes (2000), a técnica é composta por cinco etapas essenciais:

(i) definição do tema em análise;

(ii) determinação do objetivo, bem como das metas e submetas;

(iii) construção da Árvore de Decisão;

(iv) revisão da estrutura da árvore;

(v) encerramento do processo.

Dentre as vantagens associadas ao uso de Árvores de Decisão, destacam-se: a ausência de pressupostos quanto à distribuição dos dados; a flexibilidade em utilizar características ou atributos tanto qualitativos (categóricos) quanto quantitativos (numéricos); a possibilidade de construção de modelos para qualquer função, desde que o número de exemplos de treinamento seja suficiente; e o elevado nível de compreensão que esse método proporciona. Após a elaboração da árvore, torna-se fundamental proceder à sua avaliação utilizando dados que não foram empregados no treinamento. Tal abordagem permite estimar a capacidade de generalização da árvore em relação a novos conjuntos de dados e avaliar a adaptação a diferentes situações. Além disso, possibilita a mensuração da taxa de erros e acertos verificada durante a construção da árvore, conforme destacado por Brazdil (1999).

Processo de Decisão

Conforme Paula et al. (2002), o processo de construção da árvore começa com o nó raiz, que é o critério mais relevante para a classificação dos dados. Para determinar qual atributo deve ser o primeiro (ou o mais importante), utilizamos medidas matemáticas, como o ganho de informação ou o índice de Gini, descritos a seguir.

Ganho de Informação (Baseado na Entropia)

Um dos métodos mais populares para escolher um atributo é calcular o ganho de informação, que está relacionado à entropia, uma medida de desordem ou incerteza.

A entropia $H(S)$ para um conjunto de dados S é definida como:

$$H(S) = -\sum_{i=1}^{n} p_i \log_2(p_i),$$

onde p_i é a proporção de elementos da classe i no conjunto S e n é o número total de classes. Essa fórmula nos dá uma ideia de quão "impura" ou "incerta" é a distribuição dos dados.

O ganho de informação mede a redução da entropia ao dividir o conjunto de dados com base em um determinado atributo. Para um atributo A, o ganho de informação é dado por:

$$G(S, A) = H(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} H(S_v)$$

onde S_v é o subconjunto de S que tem o valor v para o atributo A e $|S_v|$ é o número de elementos em S_v, e $|S|$ é o número total de elementos em S.

O ganho de informação mede a "pureza" do conjunto resultante após a divisão. A escolha do atributo no nó interno é feita com base no maior ganho de informação.

Índice de Gini

Outro critério popular é o índice de Gini, que mede a "impureza"de uma divisão. O índice de Gini para um conjunto S é definido como:

$$Gini(S) = 1 - \sum_{i=1}^{n} p_i^2$$

Quanto menor o valor de Gini, mais pura é a divisão, ou seja, há menos diversidade nas classes após a divisão. Assim, a árvore de decisão tende a escolher o atributo que minimiza o índice de Gini. O processo de construção da árvore começa com o nó raiz, que é o critério mais relevante para a classificação dos dados. Para determinar qual atributo

deve ser o primeiro (ou o mais importante), utilizamos medidas matemáticas, como o ganho de informação ou o índice de Gini.

2.3.6 *Random Forest*

O método *Random Forest*, introduzido por Breiman (2001), consiste na utilização de um conjunto de Árvores de Decisão para executar tarefas de classificação ou regressão. O princípio central por trás desse método é a "sabedoria das multidões", que sugere que inúmeros modelos independentes (neste caso, árvores) trabalhando em conjunto tem um desempenho superior ao de um modelo individual. Em vez de criar apenas uma árvore para realizar previsões, diversas árvores são geradas, cada uma diferente da outra. A previsão final é obtida calculando a média das previsões dessas várias árvores.

Random Forest é um modelo baseado em *bagging*, que é uma técnica de aprendizado de máquina em conjunto, na qual múltiplos modelos são combinados para formar um único modelo mais robusto e confiável. O termo "*bagging*" é uma abreviação de "*bootstrap aggregating*", referindo-se às duas etapas principais desse processo.

Conforme mencionado por James (2013), o uso de *bagging* é especialmente eficaz em Árvores de Decisão. Para aplicar essa técnica ao modelo de árvore, cria-se várias árvores de decisão utilizando os conjuntos de dados de treinamento gerados pelo método *bootstrap*. Essas árvores crescem profundamente, sem restrições no seu crescimento. Como resultado, cada árvore individual apresenta alta variabilidade, mas baixo viés.

Após a construção da floresta, ela pode ser utilizada para prever a variável resposta em novas amostras. A nova observação é processada por todas as árvores na floresta, e cada uma das B árvores gera uma previsão para essa nova entrada. Denotemos a previsão feita pela árvore b para a entrada y como y^b. A previsão final do modelo é:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} y^b$$

2.3.7 Análise de Componentes Principais

A Análise de Componentes Principais ou *Principal Component Analysis* (PCA) realiza uma transformação linear nos dados de forma que as componentes mais significativas sejam alocadas nas primeiras dimensões, em eixos chamados de principais. Conforme Vasconcelos (2007), a matriz de transformação usada na PCA é composta pelos autovetores da matriz de covariância dos dados estimados. Cada coluna dessa matriz de transformação corresponde a um autovetor da matriz de covariância Σ. Esta última é uma matriz simétrica e definida positiva, que contém informações sobre as variâncias nos diferentes eixos onde os dados estão distribuídos. A matriz de covariância pode ser estimada através da

fórmula:

$$\Sigma = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu)^r(x_i - \mu)$$

onde N representa o número de amostras x_i, e μ é a média do conjunto de dados.

Os autovetores desta matriz formam uma nova base que captura a direção de maior variação dos dados. Portanto, a PCA pode ser entendida como uma mudança de base. Em essência, a PCA e a decomposição por autovalores de uma matriz são processos equivalentes, apenas abordados de perspectivas diferentes.

A aplicação da PCA a uma imagem colorida pode ser realizada em três etapas principais: primeiro, gera-se a matriz Σ com base nas operações descritas a seguir:

$$\Sigma = \text{cov}([RGB])$$

A matriz Σ obtida é uma matriz 3x3 que representa a matriz de covariância da imagem colorida. Essa matriz será utilizada para calcular a transformação que levará a imagem do espaço RGB para um novo espaço de representação, definido pela PCA. Com a matriz de covariância Σ, é possível calcular seus autovalores e autovetores, conforme representado na equação:

$$[T, aut] = \text{eig}(\Sigma)$$

A partir desta operação, obtemos as matrizes T e aut. A matriz T possui em suas linhas os autovetores da matriz de covariância, enquanto aut é uma matriz diagonal cujos elementos na diagonal correspondem aos autovalores de Σ. A operação de obtenção dos autovalores e autovetores da matriz Σ é denotada por eig.

2.3.8 Técnicas de Normalização

No aprendizado de máquina, a normalização dos dados é uma etapa crucial para garantir que os modelos aprendam de maneira eficiente e precisa. A normalização ajuda a ajustar as variáveis em escalas semelhantes, o que pode melhorar o desempenho de muitos algoritmos de ML. A seguir, estão descritas algumas das principais técnicas de normalização, com suas definições matemáticas, de acordo com Alpaydin (2020) e Singh e Singh (2020).

StandardScaler

O *StandardScaler*, ou *Z-score*, padroniza os dados de forma que eles tenham média zero e desvio padrão igual a um. Essa técnica é particularmente útil para algoritmos que assumem que os dados seguem uma distribuição normal, como Regressão Linear e Redes Neurais. A fórmula matemática para cada amostra x_i é dada por:

$$z_i = \frac{x_i - \mu}{\sigma},$$

onde z_i é o valor normalizado; x_i é o valor original; μ é a média dos dados e σ é o desvio padrão dos dados.

MinMaxScaler

O *MinMaxScaler* transforma os dados para um intervalo específico, geralmente entre 0 e 1. É útil quando as distribuições dos dados não são gaussianas e quando é necessário preservar a escala original em uma faixa limitada. A transformação é feita pela fórmula:

$$x'_i = \frac{x_i - \min(X)}{\max(X) - \min(X)},$$

onde x'_i é o valor normalizado; x_i é o valor original; $\min(X)$ e $\max(X)$ são os valores mínimos e máximos do conjunto de dados.

MaxAbsScaler

O *MaxAbsScaler* escala os dados de forma que os valores absolutos dos dados fiquem dentro do intervalo $[-1, 1]$. Ele mantém a dispersão dos dados e é particularmente útil para dados esparsos.

A fórmula é dada por:

$$x'_i = \frac{x_i}{\max(|X|)}$$

onde x'_i é o valor normalizado; x_i é o valor original e $\max(|X|)$ é o valor máximo absoluto no conjunto de dados.

RobustScaler

O *RobustScaler* utiliza a mediana e o intervalo interquartil para normalizar os dados, sendo resistente a *outliers*[1]. Isso o torna útil para dados que contenham valores extremos.

A fórmula de normalização é dada por:

$$x'_i = \frac{x_i - \overline{X}}{\mathrm{IQR}(X)},$$

onde x'_i é o valor normalizado; x_i é o valor original; $\overline{X}$ é a mediana do conjunto de dados e $\mathrm{IQR}(X)$ é o intervalo interquartil.

[1]De acordo com Alpaydin (2020), um *outlier* é uma observação que se desvia significativamente do restante dos dados. Esses dados podem distorcer as análises e influenciar negativamente os resultados dos modelos de aprendizado de máquina.

Normalizer

O *Normalizer* transforma os dados para que cada amostra individual tenha uma norma (comprimento) unitária. Ele é útil quando o foco está nas direções dos vetores de características e não em suas magnitudes. A normalização é feita usando a norma l_2:

$$x'_i = \frac{x_i}{\sqrt{\sum_{j=1}^{n} x_j^2}},$$

onde x'_i é o valor normalizado da i-ésima amostra; x_i é o valor original e soma no denominador é a norma l_2 da amostra.

2.3.9 Métricas de Avaliação

As métricas de avaliação são ferramentas essenciais para mensurar o desempenho de modelos preditivos, permitindo avaliar a precisão e a robustez das previsões realizadas (Alpaydin, 2020). Elas fornecem uma forma objetiva de comparar diferentes modelos ou abordagens, auxiliando na escolha da melhor estratégia para problemas específicos. As métricas utilizadas neste trabalho foram o Erro Percentual Absoluto Médio (*Mean Absolute Percentage Error* - MAPE), que expressa o erro em termos percentuais; o Erro Quadrático Médio (*Mean Squared Error* - MSE), que penaliza erros maiores de forma mais acentuada; e o Coeficiente de Determinação (R^2), que indica o quão bem o modelo explica a variação dos dados. A seguir, as definições de cada uma de acordo com Othmane et al. (2024).

Erro Percentual Absoluto Médio (MAPE)

O MAPE mede a magnitude do erro entre os valores previstos e os valores reais, expressando-o como uma porcentagem. A definição é dada por:

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100,$$

onde y_i são os valores reais e $\hat{y}_i$ são os valores previstos.

O MAPE é útil para interpretar o erro em termos percentuais, especialmente em previsões financeiras. No entanto, é sensível a valores muito pequenos, o que pode distorcer o erro. Para identificar possíveis problemas com essa sensibilidade, outras duas métricas foram consideradas.

Erro Quadrático Médio (MSE)

O Erro Quadrático Médio, conhecido como *Mean Squared Error* (MSE), mede a média dos quadrados dos erros, ou seja, a diferença entre os valores reais e os valores

preditos pelo modelo. O MSE é especialmente útil para capturar grandes desvios, pois, como os erros são elevados ao quadrado, penaliza de forma mais intensa erros maiores. Segue sua definição:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

onde n é o número de observações; y_i são os valores observados (dados reais) e $\hat{y}_i$ são os valores preditos pelo modelo.

O erro médio quadrático das previsões do modelo é medido em relação aos valores reais, resultando em um valor não negativo, onde $MSE = 0$ indica que o modelo fez predições perfeitas, com todos os valores preditos exatamente iguais aos valores reais. Por outro lado, valores maiores de MSE indicam que o modelo cometeu maiores erros nas previsões.

Dado que os erros são elevados ao quadrado, o método penaliza mais fortemente erros grandes do que erros pequenos, tornando-o sensível a *outliers*. No entanto, esse comportamento pode ser vantajoso em casos onde é preferível evitar grandes desvios nas previsões.

Coeficiente de Determinação (R^2)

O coeficiente de determinação (R^2) mede a proporção da variância total nos dados que é explicada pelo modelo, sendo uma forma de verificar o quão bem o modelo se ajusta aos dados observados.

O R^2 é calculado comparando o erro das previsões feitas pelo modelo com o erro das previsões que poderiam ser feitas usando simplesmente a média dos valores observados. A fórmula para o cálculo do R^2 é a seguinte:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2},$$

onde y_i são os valores observados (dados reais); $\hat{y}_i$ são os valores preditos pelo modelo e $\bar{y}$ é a média dos valores observados, dada por $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

De acordo com Müller e Guido (2016), os resultados podem ser interpretados da seguinte forma:

- $R^2 = 1$: O modelo ajusta perfeitamente os dados e explica 100% da variação nos valores observados.
- $R^2 = 0$: O modelo não consegue explicar nenhuma variação nos dados, e o uso da média dos valores observados como predição seria tão eficaz quanto o modelo.

- $R^2 < 0$: O modelo é pior do que simplesmente usar a média dos valores observados como predição. Isso sugere que o modelo está produzindo previsões muito imprecisas.

3 Metodologia

Como mencionado, este trabalho tem como objetivo prever o faturamento anual da Riot Games com o jogo League of Legends, utilizando técnicas de Estatística e Aprendizado de Máquinas. O problema consiste em criar modelos preditivos que utilizem variáveis históricas relacionadas ao jogo, como número de campeões e *skins* lançadas no ano, suas características, gênero, eventos sazonais e outras, para prever o faturamento anual da empresa.

3.1 Conjunto de Dados

A obtenção de dados detalhados sobre o faturamento da Riot Games, especificamente relacionados ao jogo League of Legends, enfrenta diversas dificuldades. Como a Riot Games é uma empresa privada, não há exigência legal para divulgar publicamente suas finanças detalhadas. Além disso, métricas estratégicas, como o número de jogadores ativos, receita de microtransações e a desempenho de eventos sazonais, são mantidas em sigilo pela empresa, tornando o acesso a essas informações mais restrito. Consequentemente, encontrar fontes diretas sobre o faturamento é um desafio significativo.

Dada essa limitação, o conjunto de dados utilizado neste estudo foi obtido de fontes abertas e confiáveis, como o site colaborativo *League of Legends Wiki*, especificamente na página localizada em League of Legends Fandom (Fandom, 2024). Este site é uma plataforma colaborativa dedicada a reunir informações detalhadas sobre o jogo e seus conteúdos, e oferece uma rica base de dados sobre aspectos fundamentais de League of Legends, especialmente em relação à *skins*, campeões e eventos.

Os dados coletados contêm informações detalhadas sobre:

- **Campeões**: O número de campeões lançados por ano, além das atualizações e reformulações de campeões existentes, que podem impactar diretamente o engajamento dos jogadores.
- ***Skins***: A quantidade de *skins* lançadas ao longo dos anos, suas variações e os tipos de *skins* disponíveis. *Skins* são uma das principais fontes de receita para a Riot Games, ao serem itens cosméticos altamente populares.
- ***Chromas***: Variações de cores adicionais para as *skins*, oferecendo mais opções de personalização aos jogadores.

- **Novos efeitos, animações e *recalls***: Algumas *skins* vêm com novos efeitos visuais, animações especiais e mudanças nas animações de *recall* (quando o personagem retorna à base), o que aumenta seu apelo.
- **Novas vozes e citações**: *Skins* ou campeões que recebem novas dublagens ou citações dentro do jogo.
- **Transformações e atualizações visuais (VU)**: Algumas *skins* e campeões passam por atualizações visuais ou transformações que melhoram significativamente sua aparência, aumentando o interesse dos jogadores em adquirir esses itens.
- **Distribuição e disponibilidade**: Informações sobre como e quando as *skins* e campeões são distribuídos, incluindo pacotes de eventos e promoções sazonais.
- **Custo e disponibilidade**: Os custos das *skins* e itens cosméticos, bem como sua disponibilidade no jogo ao longo do tempo.
- **Lançamento**: As datas de lançamento das *skins* e campeões, que são fatores críticos para identificar períodos de alta demanda e impacto no faturamento.

Essas informações são essenciais para compreender os fatores que impulsionam o faturamento da Riot Games, uma vez que novos lançamentos de *skins*, campeões e atualizações de conteúdo geralmente coincidem com aumentos significativos nas receitas. Embora os dados não incluam diretamente informações financeiras, as variáveis relacionadas ao lançamento de conteúdo cosmético e eventos servem como indicadores valiosos para prever o faturamento anual da empresa com o jogo League of Legends.

3.1.1 Dados de Faturamento

Além das informações relacionadas aos campeões e *skins*, o estudo também utilizou estimativas de faturamento anual da Riot Games com League of Legends. Como mencionado anteriormente, a empresa não divulga publicamente seus dados financeiros, logo os valores de faturamento foram obtidos a partir de discussões em fóruns e comunidades online dedicadas ao jogo. Essas fontes são geralmente baseadas em análises de mercado e estimativas de especialistas da indústria, e embora não sejam oficiais, fornecem uma referência aproximada dos ganhos da Riot Games ao longo dos anos.

Dentre os fóruns e sites utilizados para a coleta dessas estimativas, os principais são:

- **DBLTAP** (Dacanay, 2021) - Site de notícias e conteúdo especializado em jogos eletrônicos, focado principalmente em esports e cultura de games. Ele oferece notícias, análises e guias sobre diversos jogos competitivos, como League of Legends e outros.

- **Mobile Marketing Reads** (Mobile Marketing Reads, 2024) - Site de notícias e análises sobre *marketing* móvel, monetização de aplicativos e o mercado de jogos móveis. Também cobre tendências e dados relacionados a jogos populares, comportamento dos usuários e estratégias de monetização.
- **Zippia** (Zippia, 2024) - O site Zippia é uma plataforma dedicada a estudos e estatísticas sobre o mercado de trabalho nos EUA.

Os dados de faturamento obtidos a partir desses fóruns representam apenas uma estimativa e não devem ser considerados como números oficiais. No entanto, são suficientemente precisos para serem utilizados como referência no contexto deste estudo de predição de faturamento.

Tabela 3.1: Estimativas de Faturamentos Anuais de League of Legends.

Ano	**Bilhões de USD**
2015	1.63
2016	1.8
2017	2.1
2018	1.4
2019	1.5
2020	1.75
2021	1.63
2022	1.8
2023	1.5
2024	-

Os valores dispostos na Tabela 3.1 representam as estimativas de faturamento ao longo dos anos, fornecendo uma base para os modelos de predição de faturamento discutidos nas próximas seções deste trabalho.

3.2 Pré-processamento dos Dados

Os dados coletados sobre os campeões e *skins* do jogo League of Legends foram organizados em um dicionário *Python*[1], onde cada campeão possui suas respectivas *skins* e

[1]Um dicionário em *Python* é uma estrutura de dados que armazena pares de chave e valor, permitindo o acesso rápido a um valor específico a partir de sua chave. As chaves são únicas e imutáveis, como strings ou números, enquanto os valores podem ser de qualquer tipo. A sintaxe utiliza colchetes { } para criar o dicionário, com os pares separados por dois pontos. Por exemplo: {"nome": "João", "idade": 25}. Essa estrutura é amplamente utilizada para organizar dados e facilitar a recuperação eficiente de informações.

informações associadas. Esse formato estruturado permitiu que as informações fossem facilmente adaptadas com pequenas modificações, como substituições de caracteres e ajustes em valores booleanos[2] (transformados em 1 e 0), o que facilitou a importação e manipulação dos dados para a criação dos vetores de características. O resultado final desse pré-processamento está disponível nos anexos deste trabalho.

3.2.1 Descrição da Formação dos Vetores

(i) **Importação e seleção de variáveis**: Os dados de campeões e *skins* foram importados e armazenados em um dicionário contendo as principais características de cada *skin*, como nome, efeitos, animações, entre outras variáveis. Algumas variáveis consideradas irrelevantes para a predição foram descartadas, elas são: `voiceactor`, `looteligible`, `formatname`, `splashartist`, `lore`, `extras`, `formicon`, `forms`, `variant`, `earlysale`, `retired`, `removed`.

(ii) **Conversão de valores booleanos**: Informações como a presença de novos efeitos visuais ou novos *recalls* foram originalmente armazenadas como valores booleanos (`True` ou `False`). Durante o pré-processamento, essas variáveis foram convertidas para valores numéricos, onde `True` foi substituído por 1 e `False` por 0, facilitando o uso dessas informações nos modelos de aprendizado de máquina.

(iii) **Criação de dicionários para variáveis por ano**: O conjunto de dados foi estruturado para criar vetores de características por ano, de 2009 a 2024. Para cada ano, as variáveis relevantes foram coletadas e organizadas. As principais variáveis independentes incluem:

- **Campeões lançados por ano**: O número de novos campeões lançados a cada ano.
- **Gênero dos campeões**: Contagem de campeões masculinos, femininos ou sem gênero específico por ano.
- **Classe dos campeões**: Distribuição de classes (ex.: lutadores, magos, atiradores) para os campeões lançados no respectivo ano.
- **Posição dos campeões**: A função ou posição principal dos campeões no jogo (ex.: topo, meio, suporte).
- ***Skins* lançadas por ano**: Quantidade de *skins* lançadas em cada ano, incluindo variações de *chromas*.

[2]Valores booleanos são um tipo de dado que assume apenas dois valores: verdadeiro (*True*) ou falso (*False*), amplamente usados em lógica computacional para controlar fluxos de execução e realizar comparações lógicas (Cormen, 2017).

- ***Labels* e disponibilidade**: As *skins* foram categorizadas por disponibilidade, considerando se foram lançadas em eventos especiais ou em outros contextos limitados, além de atualizações como novas animações, efeitos, vozes e outras.
- ***Sets* e temas agrupados**: Alguns temas de *skins* foram agrupados em 19 grupos por similaridade, o que resultou em 150 conjuntos únicos ao longo dos anos.

(iv) **Geração de vetores de variáveis independentes**: Após a coleta das variáveis relevantes para cada ano, os dados foram combinados para formar vetores de características que representassem cada ano entre 2009 e 2024. Cada vetor resultante possui dimensão 179, incorporando todas as variáveis mencionadas anteriormente, incluindo a contagem de campeões, gênero, classe, posição, número de *skins*, *labels* e disponibilidade, além dos conjuntos temáticos. Esses vetores serviram como entrada para os modelos de aprendizado de máquina.

(v) **Considerações adicionais**: O custo de cada *skin* em Riot Points (RP) não foi incluído na formação dos vetores. Essa decisão foi tomada devido à alta variabilidade dos preços das *skins*, que poderia introduzir um desbalanceamento entre as variáveis. *Skins* de categorias muito diferentes (por exemplo, *skins* épicas e *skins* comuns) apresentam grandes discrepâncias de custo que, se incluídas nos vetores, poderiam dominar o modelo e ofuscar a importância de outras variáveis. Além disso, como o foco está em prever o faturamento global baseado em lançamentos e não no comportamento individual dos preços, a contagem de *skins* e sua disponibilidade foram consideradas mais representativas para a formação dos vetores.

Este processo de pré-processamento garantiu que os dados estivessem estruturados de forma coerente e padronizada, prontos para serem usados nos modelos de previsão de faturamento descritos nas próximas seções deste trabalho.

3.3 Modelos de Aprendizado de Máquina e Técnicas Utilizadas

Foram utilizados três modelos de aprendizado de máquina supervisionado neste estudo: Regressão Linear, *Random Forest* e Árvore de Decisão. Além disso, foi aplicada a técnica de redução de dimensionalidade por meio da Análise de Componentes Principais (PCA), com o objetivo de simplificar o conjunto de variáveis e potencializar o desempenho dos modelos. Esses métodos foram selecionados por sua capacidade de capturar padrões nos dados e suas diferentes abordagens para modelagem preditiva. A seguir, são descritos os modelos e a técnica utilizada.

- **Regressão Linear**: A regressão linear foi utilizada como modelo inicial por sua simplicidade e interpretabilidade. Ela assume uma relação linear entre as variáveis independentes e o faturamento anual, ajustando coeficientes para minimizar o erro entre as previsões e os valores reais. Embora seja fácil de interpretar, a regressão linear pode não capturar relações não lineares presentes nos dados.

- **Árvore de Decisão**: O modelo de árvore de decisão foi utilizado por sua capacidade de capturar relações não lineares e interações entre variáveis. A árvore de decisão divide os dados em subconjuntos baseados nas variáveis mais importantes, criando uma estrutura hierárquica de decisões. Apesar de sua eficiência, este modelo pode ser propenso a *overfitting*[3] se não for regularizado.

- ***Random Forest***: Para mitigar o risco de *overfitting* das árvores de decisão, o modelo de *Random Forest* foi empregado. Ele combina várias árvores de decisão geradas a partir de diferentes subconjuntos dos dados, resultando em um modelo mais robusto e generalizável. Essa abordagem reduz a variância e melhora a performance em dados de teste.

- **Análise de Componentes Principais (PCA)**: Embora não seja um modelo preditivo por si só, o PCA foi utilizado para redução de dimensionalidade. Esta técnica transformou as variáveis originais em componentes principais, permitindo que os modelos fossem treinados com um número menor de variáveis, sem perda significativa de informação, melhorando assim a eficiência e a interpretabilidade dos modelos.

Esses modelos foram escolhidos para explorar diferentes relações, lineares e não lineares, entre as variáveis independentes/explicativas e o faturamento da Riot Games.

3.4 Validação e Avaliação dos Modelos

Para a metodologia proposta, foi criado um vetor com as variáveis independentes de cada ano no período de 2015 a 2024. Estas variáveis incluem dados como número de campeões e *skins* lançados em cada ano, atualizações, eventos sazonais, entre outros mencionados anteriormente.

Devido à natureza do problema e à baixa quantidade de amostras disponíveis (9 anos de dados no total), os dados foram divididos em dois subconjuntos:

[3] *Overfitting* é o fenômeno em que um modelo de aprendizado de máquina se ajusta excessivamente aos dados de treinamento, capturando ruídos e padrões irrelevantes, o que resulta em um excelente desempenho no conjunto de treinamento, mas em uma baixa capacidade de generalização para novos dados, prejudicando as previsões (Goodfellow et al., 2016).

- **Conjunto de treinamento**: Dados dos anos de 2015 a 2021. Esse período foi utilizado para treinar os modelos de aprendizado de máquinas.
- **Conjunto de teste**: Dados dos anos de 2022 e 2023. Este conjunto foi reservado para testar a performance dos modelos e avaliar sua capacidade de generalização para prever os valores futuros.

A escolha por esta divisão se justifica pela natureza temporal dos dados e pela quantidade limitada de amostras. Com um total de apenas 9 anos, a validação cruzada[4] não foi utilizada, pois essa técnica geralmente é mais eficaz em conjuntos de dados maiores, onde há uma quantidade suficiente de amostras para gerar múltiplas partições de treino e validação sem comprometer a quantidade de dados de cada conjunto. Utilizar validação cruzada com um número tão limitado de amostras poderia resultar em partições com pouca representatividade, prejudicando a capacidade dos modelos de capturar padrões robustos nos dados.

Assim, optou-se por um método de validação simples, onde o conjunto de teste (2022-2023) foi mantido separado, garantindo que o modelo fosse avaliado em dados nunca antes vistos, o que proporciona uma estimativa mais confiável de sua performance futura. As métricas escolhidas para a avaliação dos modelos foram o MAPE, MSE e R^2, definidos no Capítulo anterior. Essas métricas foram escolhidas por sua eficácia em medir o desempenho de modelos de regressão (Pedregosa, 2011), especialmente no contexto de predição de séries temporais (Müller; Guido, 2016).

3.5 Ferramentas e Bibliotecas

O desenvolvimento deste trabalho foi realizado utilizando a linguagem de programação *Python*, que oferece uma ampla gama de bibliotecas adequadas para tarefas de aprendizado de máquina, análise de dados e visualização. Para a implementação, o ambiente de desenvolvimento utilizado foi o Google Colab, que oferece recursos computacionais em nuvem, facilitando o processamento de grandes volumes de dados, quando necessário, e a execução de modelos de aprendizado de máquina.

As principais bibliotecas amplamente empregadas em estudos de Aprendizado de Máquina (Müller; Guido, 2016; Unpingco, 2022), e também neste trabalho, foram:

- **Pandas** (`import pandas as pd`): Utilizada para manipulação e análise de dados, incluindo operações de leitura de dados, organização em *dataframes* e transformações

[4] A validação cruzada é uma técnica usada para avaliar o desempenho de um modelo de maneira mais confiável. Em vez de simplesmente dividir os dados em um conjunto de treino e teste, ela divide os dados em várias partes e testa o modelo em diferentes combinações, garantindo uma avaliação mais precisa e robusta da capacidade de generalização do modelo Müller e Guido (2016).

necessárias para o pré-processamento.

- **NumPy** (`import numpy as np`): Biblioteca fundamental para a manipulação de *arrays* e operações matemáticas de alto desempenho.
- **Scikit-learn**:
  - **Pré-processamento**: As funções `StandardScaler`, `MinMaxScaler`, `MaxAbsScaler`, `RobustScaler` e `Normalizer` foram utilizadas para normalizar e padronizar as variáveis independentes, garantindo que todas estivessem na mesma escala, conforme as necessidades dos dados.
  - **Modelos**: Diversos modelos de aprendizado de máquina foram implementados a partir desta biblioteca, incluindo:
    - Regressão Linear (`from sklearn.linear_model import LinearRegression`),
    - Árvore de Decisão (`from sklearn.tree import DecisionTreeRegressor`),
    - *Random Forest* (`from sklearn.ensemble import RandomForestRegressor`),
  - **Métricas**: Para avaliar o desempenho dos modelos, foram utilizadas as funções `mean_squared_error`, `r2_score` e `mean_absolute_percentage_error`, que calculam, respectivamente, o MSE, o R^2 e o MAPE.
  - **Validação e Divisão de Dados**: A função `train_test_split` foi utilizada para dividir os dados em conjuntos de treino e teste.
  - **PCA** (`from sklearn.decomposition import PCA`): O algoritmo de Análise de Componentes Principais foi utilizado para redução de dimensionalidade em alguns experimentos, permitindo reduzir a quantidade de variáveis independentes e melhorar o desempenho dos modelos.
- **Seaborn** (`import seaborn as sns`) e **Matplotlib** (`import matplotlib.pyplot as plt`): Estas bibliotecas foram utilizadas para a criação de gráficos e visualizações, facilitando a análise exploratória de dados e a interpretação dos resultados dos modelos.

O Google Colab, por sua vez, proporcionou um ambiente de desenvolvimento interativo, permitindo a execução de código em tempo real, além de oferecer suporte a GPUs para acelerar o treinamento dos modelos de aprendizado de máquina, quando necessário.

4 Desenvolvimento e Resultados

Este capítulo apresenta o processo de desenvolvimento e os resultados obtidos a partir da implementação dos métodos de predição de faturamento da Riot Games, utilizando dados históricos relacionados ao jogo League of Legends. Inicialmente, será descrito o pré-processamento dos dados, seguido da construção dos vetores de variáveis independentes e da preparação para a aplicação dos modelos preditivos. Em seguida, será apresentada a avaliação dos modelos de aprendizado de máquina utilizados, com base em métricas de desempenho como o *Mean Squared Error* (MSE) e o coeficiente de determinação (R^2). Os resultados serão discutidos em detalhes, considerando as diferentes abordagens e ajustes realizados ao longo do processo, com o intuito de otimizar as previsões de faturamento.

4.1 Coleta e Preparação dos Dados

No desenvolvimento do trabalho, os dados obtidos da plataforma *League of Legends Fandom* (Fandom, 2024) foram convertidos para um dicionário em *Python*, como mencionado anteriormente, ficando estruturados como no exemplo exibido no Código 4.1. Esse processo envolveu a substituição de alguns caracteres específicos para adequar o formato original ao padrão de um dicionário, facilitando a manipulação e o tratamento das informações subsequentes.

Código 4.1: Dados no formato dicionário, na linguagem de programação *Python*.

```
1
2  dados = {
3    "Aatrox" : {
4      "id" : 266,
5      "skins" : {
6        "Original" : {
7          "id" : 0,
8          "availability" : "Available",
9          "cost" : 880,
10         "release" : "2013-06-12",
11         "voiceactor" : {"Ramon Tikaram"},
12         "splashartist" : {"Victor '3rdColossus' Maury"},
13         "lore" : "Once honored defenders of Shurima against the
               Void, Aatrox and his brethren would eventually become an
                even greater threat to Runeterra, and were defeated
               only by cunning mortal sorcery. But after centuries of
```

```
            imprisonment, Aatrox was the first to find freedom once
            more, corrupting and transforming those foolish enough
            to try and wield the magical weapon that contained his
            essence. Now, with stolen flesh, he walks Runeterra in a
             brutal approximation of his previous form, seeking an
            apocalyptic and long overdue vengeance."
        },
        "Justicar" : {
          "id" : 1,
          "availability" : "Available",
          "looteligible" : True,
          "cost" : 975,
          "release" : "2013-06-12",
          "set" : {"Justicar"},
          "neweffects" : True,
          "newanimations" : True,
          "newrecall" : True,
          "filter" : True,
          "voiceactor" : {"Ramon Tikaram"},
          "splashartist" : {"Pan Chengwei"},
          "lore" : "Chief among the Arclight stand the Justicars,
             sentinels who serve as living embodiments of justice and
              order. Aatrox embodies power, for it is his martial
             prowess that holds the chaos back."
        },
        .
        .
        .
    }
  },
  .
  .
  .
}
```

4.2 Pré-processamento dos Dados

Em seguida, foi implementado o pré-processamento dos dados, conforme descrito no capítulo anterior, com o objetivo de gerar vetores de variáveis independentes (também conhecidos como vetores de características, *features* em inglês) para cada ano analisado. Esses vetores contêm informações relevantes sobre os campeões e *skins* lançados, bem como

eventos e outras variáveis categóricas associadas, como a disponibilidade de *skins* e temas recorrentes.

O código responsável pelo pré-processamento organiza essas informações de maneira sistemática, transformando os dados brutos em um formato numérico estruturado. Variáveis booleanas foram convertidas para valores binários (1 ou 0), enquanto categorias, como a disponibilidade das *skins*, foram ponderadas com base em sua relevância. Além disso, o agrupamento de *skins* por similaridade temática, por meio de dicionários auxiliares, permitiu a criação de variáveis adicionais que refletissem o uso recorrente de certos conjuntos visuais ao longo dos anos.

Ao final, o processamento resultou em vetores anuais contendo uma síntese quantitativa de 179 características, que serviram como entradas para os modelos preditivos, alinhados com os objetivos do projeto. Os vetores gerados no processo de pré-processamento descrito possuem o formato do exemplo detalhado no Código 4.2.

Código 4.2: Vetor de variáveis independentes, do ano de 2022, gerado durante o pré-processamento dos dados.

```
1
2  vetores = {...,
3
4      '2022': [5, 1, 4, 0, 1, 0, 2, 1, 1, 1, 1, 2, 1, 0, 58, 1037,
           163, 163, 163, 32, 10, 0, 4, 0, 35, 118, 16, 18, 17, 2, 12,
           0, 0, 4, 1, 0, 0, 2, 4, 0, 0, 12, 0, 6, 3, 7, 13, 6, 0, 0,
           1, 0, 6, 0, 0, 0, 0, 0, 0, 4, 10, 5, 0, 0, 0, 0, 0, 1, 1, 0,
            9, 0, 1, 0, 11, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1,
           0, 0, 1, 0, 0, 0, 0, 0, 0, 6, 0, 0, 0, 0, 0, 0, 0, 0, 0, 8,
           3, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 6, 0,
           0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
           0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 3, 0, 0, 4,
           0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
5
6      ...
7  }
```

Cada coordenada dos vetores representam um quantitativo de uma característica específica dos campeões e *skins* lançados por ano, como descrito a seguir:

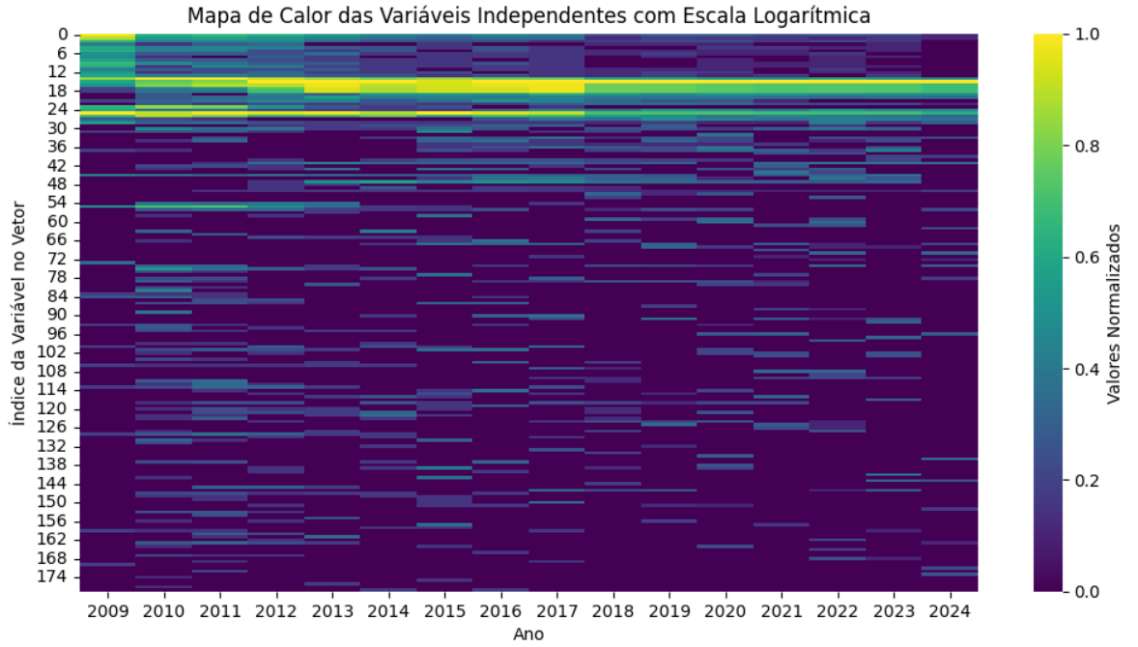
- **Coordenada 1**: Número de novos campeões lançados no ano;
- **Coordenadas 2 e 3**: Número de campeões dos gêneros masculino e feminino, respectivamente, lançados no ano;
- **Coordenadas 4 a 9**: Número de campeões das classes *Assassin*, *Fighter*, *Mage*, *Marksman*, *Support*, *Tank*, respectivamente, lançados no ano;

- **Coordenadas 10 a 14**: Número de campeões das posições *TOP*, *JUNGLE*, *ADC*, *SUP*, *MID*, respectivamente, lançados no ano;
- **Coordenada 15**: Número de novas *skins* lançadas no ano;
- **Coordenadas 16 a 25**: Alterações/Atualizações feitas no ano para todos os campeões e *skins* disponíveis, sendo: *chromas*, *neweffects*, *newanimations*, *newrecall*, *filter*, *newvoice*, *newquotes*, *transforming*, *vu*, *distribution*, respectivamente;
- **Coordenadas 26 a 29**: Contagem de campeões/*skins* disponíveis no ano conforme sua disponibilidade, sendo: *Available*, *Legacy*, *Rare*, *Limited*, respectivamente;
- **Coordenadas 30 a 179**: Contagem de eventos realizados no ano para cada campeão/*skin*. Cada coordenada corresponde a um evento ou grupo de eventos similares.

Os dados coletados sobre as estimativas de faturamento, apresentados na Tabela 3.1, também foram organizados no formato de dicionário, com cada ano como chave e seu respectivo valor de faturamento em bilhões de dólares. Essas estimativas representam as saídas esperadas dos modelos preditivos. Utilizando esses dados como referência, os modelos de aprendizado de máquina serão treinados para prever o faturamento futuro da Riot Games, com base nas variáveis independentes associadas aos lançamentos e eventos.

A seguir, os vetores de variáveis independentes foram normalizados e representados em escala logarítmica no mapa de calor mostrado na Figura 4.1. Para isso, utilizou-se um processo de normalização que ajusta os dados a um intervalo entre 0 e 1, facilitando a visualização comparativa entre os diferentes anos e variáveis analisadas. Foi aplicada uma transformação logarítmica para reduzir o impacto das grandes variações entre variáveis com escalas significativamente diferentes das demais, como os lançamentos de *skins* e *chromas*, que costumam ocorrer em maiores volumes.

Figura 4.1: Mapa de calor das variáveis independentes por ano, normalizadas e com escala logarítmica. O gráfico destaca a intensidade das variáveis analisadas de 2009 a 2024, incluindo lançamentos de campeões e *skins*, com uma tendência decrescente nos lançamentos de novos campeões nos últimos anos.



O mapa de calor (Figura 4.1) proporciona uma visão geral das variáveis analisadas ao longo do tempo, de 2009 a 2024. Cada linha do gráfico representa uma variável, e cada coluna corresponde a um ano, com as cores indicando a intensidade de cada variável em um determinado ano. A cor mais clara (amarelo) reflete valores mais altos, enquanto os tons mais escuros (roxo) indicam valores mais baixos.

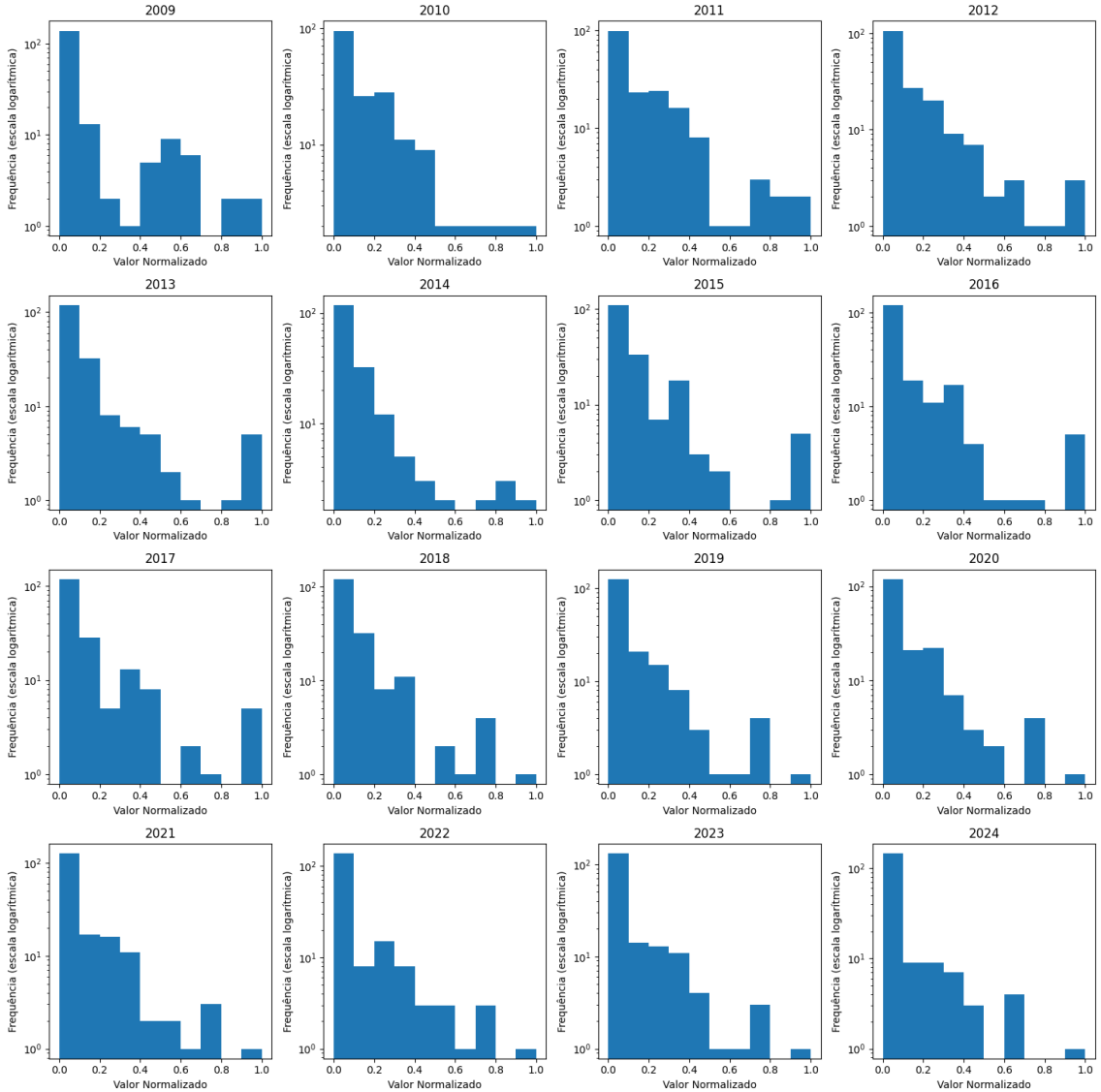
Ao analisar o gráfico, é possível observar que os lançamentos de *skins* e *chromas*, por exemplo, representados pelas linhas de índices 15 e 16, têm uma presença bastante expressiva ao longo dos anos, sendo mais intensos especialmente entre 2013 e 2017. Isso reflete um período de forte investimento nesses elementos do jogo. Em contraste, outras variáveis, como eventos e alterações relacionadas a campeões, apresentam uma distribuição mais esparsa, sugerindo que ocorrem em menor frequência ou com variação significativa ao longo dos anos.

Além disso, o gráfico revela uma tendência decrescente no número de lançamentos de novos campeões ao longo dos anos (coordenada 1 ou índice 0). Nos primeiros anos analisados, é possível observar uma maior frequência de lançamentos, especialmente em torno de 2011 a 2014. No entanto, a partir de 2015, essa quantidade começa a diminuir de forma gradual, com anos mais recentes, como 2020 e 2021, apresentando um número

consideravelmente menor de campeões novos.

Analisando os histogramas com os valores das variáveis independentes normalizadas, representados na Figura 4.2, confirma-se a tendência decrescente nos lançamentos de novos campeões e estabilização nos lançamentos de *skins* ao longo dos anos. Entre 2009 e 2017, observamos um crescimento de densidade de variáveis normalizadas concentradas em valores mais altos, o que reflete a crescente frequência de lançamentos e atualizações. Essa intensidade diminui gradualmente a partir de 2017, como ilustrado pelos histogramas, que apresentam certa estabilidade e uma distribuição mais espalhada e menos concentrada nos valores mais elevados.

Figura 4.2: Histogramas das variáveis independentes normalizadas por ano (2009-2024) em escala logarítmica.



Essa mudança pode estar associada a uma estratégia da Riot Games de equilibrar a criação de novos conteúdos com refinamentos e atualizações pontuais. Embora o período entre 2013 e 2017 tenha sido marcado por um crescimento expressivo nos lançamentos de *skins* e campeões, os histogramas mostram que a partir de 2017 há uma tendência de estabilização. Isso pode indicar que a empresa passou a concentrar seus esforços em manter o equilíbrio e a qualidade do conteúdo existente, ao invés de introduzir grandes volumes de novos lançamentos. Assim, a análise dos histogramas sugere que, ao longo do tempo, houve uma adaptação na abordagem de desenvolvimento, com foco em ajustes finos e manutenção do interesse dos jogadores, sem perder de vista a criação de novos conteúdos, mas de forma mais estratégica.

4.3 Treinamento e Testes

Para a fase de treinamento e teste foram realizados cinco experimentos utilizando as três técnicas preditivas introduzidas nos Capítulos anteriores: Regressão Linear, *Random Forest* e Árvore de Decisão. Além disso, foi aplicada a técnica de Análise de Componentes Principais (PCA) para a redução de dimensionalidade das variáveis, com o objetivo de otimizar a performance dos modelos sem comprometer a integridade das informações.

Embora tenhamos informações disponíveis para construir os vetores de características desde 2009, os dados de faturamento estão disponíveis apenas a partir de 2015, conforme mencionado no Capítulo anterior e detalhado na Tabela 3.1. Dessa forma, os dados foram divididos da seguinte maneira: os anos de 2015 a 2021 foram utilizados para o treinamento dos modelos, enquanto os anos de 2022 e 2023 foram reservados para a fase de teste. O ano de 2024 foi destinado à predição, uma vez que ainda não há informações de faturamento disponíveis para esse período. Essa separação temporal é essencial para avaliar a capacidade dos modelos em generalizar suas previsões e garantir uma análise robusta.

Em todos os experimentos, foi utilizado o parâmetro `random_state`[1] igual a 5 para garantir a reprodutibilidade dos resultados. A escolha desse número não tem nenhum significado específico, sendo utilizada apenas para manter a consistência ao longo dos experimentos e possibilitar comparações justas entre as diferentes abordagens.

O desempenho de cada modelo foi medido por meio de métricas como o Erro Quadrático Médio (MSE) e o Coeficiente de Determinação (R^2), com o objetivo de avaliar a precisão das previsões e a eficácia dos modelos em identificar os fatores mais influentes no faturamento da Riot Games. Nas seções seguintes, os resultados de cada experimento

[1]Em *Python*, o parâmetro `random_state` é utilizado em várias funções de bibliotecas para garantir a reprodutibilidade dos resultados. Ao definir um valor específico para `random_state`, o comportamento aleatório é fixado, permitindo que os mesmos resultados sejam obtidos em execuções repetidas, essencial para comparações consistentes entre experimentos.

serão apresentados, destacando as vantagens e limitações de cada abordagem.

4.3.1 Experimento 1: Predição sem Normalização dos Vetores

O primeiro experimento foi realizado sem a normalização dos vetores de características. Essa abordagem foi adotada para avaliar o desempenho dos modelos ao trabalhar com os dados em seu formato original, sem ajustes nas escalas das diferentes variáveis. Embora a normalização seja uma prática comum para melhorar o desempenho em modelos de aprendizado de máquina, especialmente em algoritmos sensíveis à magnitude das variáveis, como o *Random Forest*, a ausência dessa etapa permite que tenhamos uma linha de base sobre o impacto dessa transformação.

Foram aplicados três modelos preditivos para comparação: Regressão Linear, Árvore de Decisão e *Random Forest*. Os dados de treinamento englobam os anos de 2015 a 2021, enquanto os dados de teste abrangem os anos de 2022 e 2023. O ano de 2024 foi reservado para a fase de predição, uma vez que ainda não há dados de faturamento disponíveis para esse período. Seguem os resultados de MSE, R^2 e MAPE para as previsões na Tabela 4.1.

Tabela 4.1: MSE, R^2 e MAPE para os modelos de Regressão Linear, Árvore de Decisão e *Random Forest*

Modelo	**MSE**	R^2	**MAPE**
Regressão Linear	0.211	-8.362	0.2276
Árvore de Decisão	0.023	-0.018	0.090
Random Forest	0.021	0.083	0.087

A Regressão Linear, embora seja um modelo simples e interpretável, apresentou um desempenho insatisfatório neste experimento. O valor de MSE foi relativamente alto (0.211), o R^2 foi negativo (-8.362), e o MAPE alcançou 22.76%, indicando que o modelo não conseguiu capturar adequadamente as tendências do faturamento nos anos de teste. Essa falha pode ser atribuída à incapacidade da Regressão Linear de lidar com padrões não lineares presentes nos dados e à ausência de normalização, o que amplificou as discrepâncias entre variáveis de diferentes escalas.

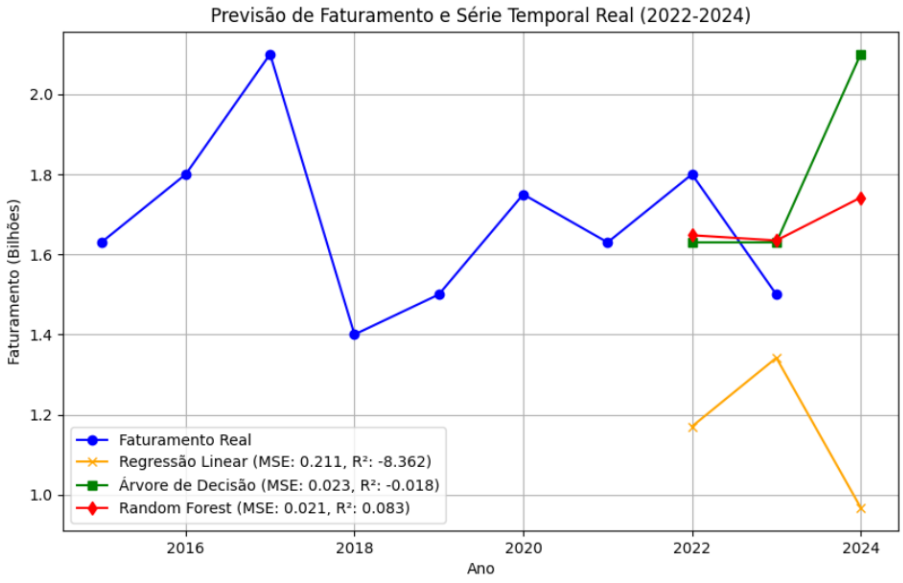
O modelo de Árvore de Decisão apresentou um MSE bem menor (0.023) e um MAPE de 9.06%, indicando que o modelo conseguiu ajustar-se melhor aos dados em comparação à Regressão Linear. No entanto, o R^2 também foi negativo (-0.018), sugerindo que o modelo ainda não conseguiu generalizar adequadamente nos anos de teste, mesmo captando melhor as variações nos dados.

Por fim, o *Random Forest* apresentou o menor MSE (0.021) e foi o único modelo com um R^2 positivo (0.083). O MAPE também foi o menor entre os três modelos (8.72%), o

que demonstra que, mesmo sem normalização, o *Random Forest* conseguiu capturar parte das dinâmicas complexas dos dados e fornecer previsões mais estáveis em comparação aos outros modelos.

Na Figura 4.3, observamos a série temporal do faturamento real (em azul) comparada com as previsões dos três modelos nos anos de teste (2022 e 2023), além da predição para 2024. A linha azul, representando os dados reais, mostra uma trajetória de crescimento variável até 2023.

Figura 4.3: Comparação entre o faturamento real e as previsões geradas pelos modelos de Regressão Linear, Árvore de Decisão e *Random Forest* nos anos de teste (2022-2023) e a predição para 2024.



As previsões da Regressão Linear (em laranja) apresentam uma queda abrupta após 2023, evidenciando a limitação do modelo em capturar as dinâmicas complexas do faturamento ao longo do tempo. Já a Árvore de Decisão (em verde) mostra uma previsão mais consistente, embora ainda insuficiente para refletir o crescimento real da empresa. O *Random Forest* (em vermelho) apresentou previsões mais estáveis e próximas dos valores reais, sendo o modelo com o melhor desempenho entre os três.

Este primeiro experimento demonstrou que a falta de normalização impactou negativamente a performance da Regressão Linear, mas teve um efeito menor sobre a Árvore de Decisão e o *Random Forest*. No entanto, os resultados sugerem que a normalização dos vetores de características pode melhorar a precisão das previsões, o que será testado no próximo experimento.

4.3.2 Experimento 2: Predição com Diferentes Técnicas de Normalização

No segundo experimento, foram aplicadas diversas técnicas de normalização aos vetores de características para avaliar o impacto dessa transformação no desempenho dos modelos. As técnicas utilizadas incluem o *StandardScaler*, *MinMaxScaler*, *MaxAbsScaler*, *RobustScaler* e *Normalizer*. A normalização foi implementada para ajustar os valores das variáveis em intervalos específicos, minimizando o impacto das discrepâncias de escala entre elas. A expectativa é que a normalização beneficie principalmente modelos como a Regressão Linear e o *Random Forest*, que são sensíveis à magnitude das variáveis.

Foram utilizados os mesmos três modelos preditivos do Experimento 1: Regressão Linear, Árvore de Decisão e *Random Forest*. Os dados de treinamento continuam a englobar o período de 2015 a 2021, enquanto os dados de teste abrangem os anos de 2022 e 2023, com a predição estendida até 2024. O objetivo deste experimento foi verificar se diferentes técnicas de normalização melhorariam o desempenho dos modelos em relação ao Experimento 1, onde os dados não foram normalizados.

Na Tabela 4.2 estão os resultados de MSE, R^2 e MAPE para cada técnica de normalização e modelo preditivo:

Tabela 4.2: MSE, R^2 e MAPE para os modelos de Regressão Linear, Árvore de Decisão e *Random Forest* utilizando diferentes técnicas de normalização.

Técnica	**Modelo**	**MSE**	R^2	**MAPE**
StandardScaler	**Regressão Linear**	0.186	-7.260	0.206
	Árvore de Decisão	0.023	-0.018	0.091
	Random Forest	0.022	0.017	0.091
MinMaxScaler	**Regressão Linear**	0.132	-4.857	0.192
	Árvore de Decisão	0.023	-0.018	0.091
	Random Forest	0.021	0.083	0.087
MaxAbsScaler	**Regressão Linear**	0.056	-1.477	0.138
	Árvore de Decisão	0.023	-0.018	0.091
	Random Forest	0.021	0.077	0.087
RobustScaler	**Regressão Linear**	0.070	-2.118	0.105
	Árvore de Decisão	0.023	-0.018	0.091
	Random Forest	0.021	0.083	0.087
Normalizer	**Regressão Linear**	0.037	-0.647	0.114
	Árvore de Decisão	0.023	-0.018	0.091
	Random Forest	0.021	0.048	0.089

Note que a normalização trouxe melhorias expressivas para a Regressão Linear, com o *Normalizer* sendo a técnica mais eficaz, resultando em um MSE de 0.037, R^2 de -

0.647 e MAPE de 0.114. Embora o R^2 ainda seja negativo, o desempenho do modelo foi consideravelmente melhor após a normalização em comparação com o Experimento 1. Para a Árvore de Decisão, no entanto, a normalização não gerou grandes benefícios. O MSE manteve-se estável em torno de 0.023 para todas as técnicas, e o R^2 permaneceu negativo, evidenciando que esse modelo não se beneficia significativamente da normalização devido à sua natureza baseada em regras, que não é sensível à escala das variáveis. Já para o *Random Forest*, houve uma leve melhoria, com o *MinMaxScaler* apresentando os melhores resultados, com MSE de 0.021, R^2 de 0.083 e MAPE de 0.087. A normalização permitiu que o modelo capturasse melhor as variações nos dados, embora os ganhos tenham sido modestos em comparação com os outros modelos.

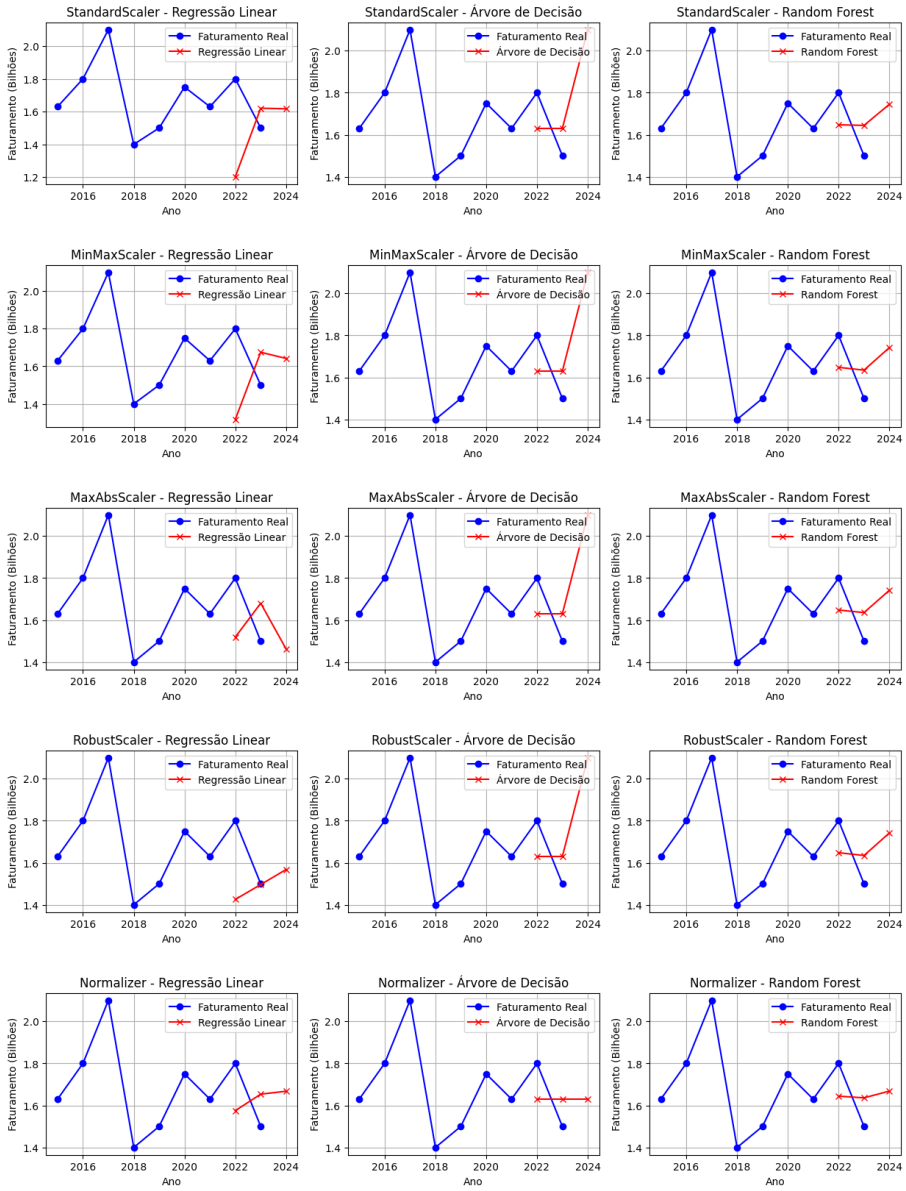
Na Figura 4.4, é possível observar a série temporal do faturamento real (em azul) comparada com as previsões dos três modelos para os anos de teste (2022-2023) e a predição para 2024, utilizando as diferentes técnicas de normalização. A Regressão Linear mostrou-se significativamente mais alinhada com a série temporal real quando normalizada, especialmente com a técnica *Normalizer*, reforçando o mencionado no parágrafo anterior. A Árvore de Decisão, por outro lado, apresentou previsões semelhantes independentemente da técnica de normalização aplicada, confirmando que a transformação não impacta tanto o modelo. O *Random Forest* também se beneficiou da normalização, especialmente com o *MinMaxScaler*, oferecendo previsões mais consistentes para os anos de 2022 a 2024. Entretanto, ainda se observa dificuldade dos modelos em capturar a variação acentuada no crescimento entre 2023 e 2024, sugerindo que ajustes de hiperparâmetros ou o uso de técnicas mais avançadas podem ser necessários.

Enfim, este segundo experimento demonstrou que a normalização é benéfica, como esperado. Com base nos resultados apresentados, nos experimentos seguintes as normas que serão usadas são:

- **Regressão Linear**: *Normalizer*
- **Árvore de Decisão e *Random Forest***: *MinMaxScaler*

No próximo experimento, será aplicada uma busca por parâmetros (*Grid Search*) (Müller; Guido, 2016) com o objetivo de otimizar o desempenho dos modelos. O *Grid Search* é uma técnica que explora de forma exaustiva todas as combinações possíveis de hiperparâmetros definidos pelo usuário para um modelo, selecionando os valores que proporcionam o melhor desempenho com base em uma métrica específica, como o MSE ou R^2. Essa abordagem será crucial para ajustar os modelos de Regressão Linear, Árvore de Decisão e *Random Forest*, considerando as técnicas de normalização selecionadas.

Figura 4.4: Comparação entre o faturamento real e as previsões geradas pelos modelos de Regressão Linear, Árvore de Decisão e *Random Forest* utilizando diferentes técnicas de normalização.



4.3.3 Experimento 3: Busca por Hiperparâmetros (Grid Search)

Neste terceiro experimento, foi realizada uma busca exaustiva por hiperparâmetros (*Grid Search*) para otimizar o desempenho dos modelos de Árvore de Decisão e *Random Forest*. O objetivo foi ajustar as configurações de cada modelo para melhorar a capaci-

dade preditiva. A Regressão Linear, por sua simplicidade e ausência de hiperparâmetros significativos, não participou desta busca.

É importante destacar que, neste experimento, a busca por hiperparâmetros foi realizada sem a aplicação de validação cruzada, ou seja, o *Grid Search* ajustou e avaliou os modelos nos mesmos dados de treinamento. Esta abordagem foi adotada devido à quantidade limitada de amostras disponíveis, com o intuito de preservar o maior número de dados possíveis para o teste e predição.

Os hiperparâmetros investigados para cada modelo são apresentados na Tabela 4.3, juntamente com seus significados:

Tabela 4.3: Hiperparâmetros investigados para Árvore de Decisão e *Random Forest*.

Modelo	**Hiperparâmetro**	**Valores Considerados**
Árvore de Decisão	*max_ depth*	3, 5, 10, 20, None
	min_ samples_ split	2, 5, 10, 20
	min_ samples_ leaf	1, 2, 4, 8
Random Forest	*n_ estimators*	50, 100, 200
	max_ depth	None, 10, 20, 30
	min_ samples_ split	2, 5, 10, 20
	min_ samples_ leaf	1, 2, 4, 5

Os hiperparâmetros são definidos da seguinte forma:

- ***n_ estimators***: Número de árvores no modelo de *Random Forest*.
- ***max_ depth***: Profundidade máxima da árvore, que controla o quão detalhada será a árvore de decisão.
- ***min_ samples_ split***: Número mínimo de amostras exigido para dividir um nó.
- ***min_ samples_ leaf***: Número mínimo de amostras em um nó folha.

Foram realizados 80 experimentos para o modelo de Árvore de Decisão (combinando 5 valores de *max_ depth*, 4 valores de *min_ samples_ split* e 4 valores de *min_ samples_ leaf*) e 192 experimentos para o modelo *Random Forest* (combinando 3 valores de *n_ estimators*, 4 valores de *max_ depth*, 4 valores de *min_ samples_ split* e 4 valores de *min_ samples_ leaf*).

Após a aplicação do *Grid Search*, foram obtidos os melhores hiperparâmetros para cada modelo, conforme mostrado na Tabela 4.4.

Tabela 4.4: Melhores Hiperparâmetros encontrados para Árvore de Decisão e *Random Forest*.

Hiperparâmetro	**Árvore de Decisão**	***Random Forest***
n_estimators	-	50
max_depth	3	None
min_samples_leaf	1	1
min_samples_split	2	2

Com esses hiperparâmetros ajustados, os modelos foram avaliados com os dados de teste dos anos de 2022 e 2023, além de fazerem previsões para o ano de 2024. A seguir, são apresentados os resultados de MSE, R^2 e MAPE para cada modelo:

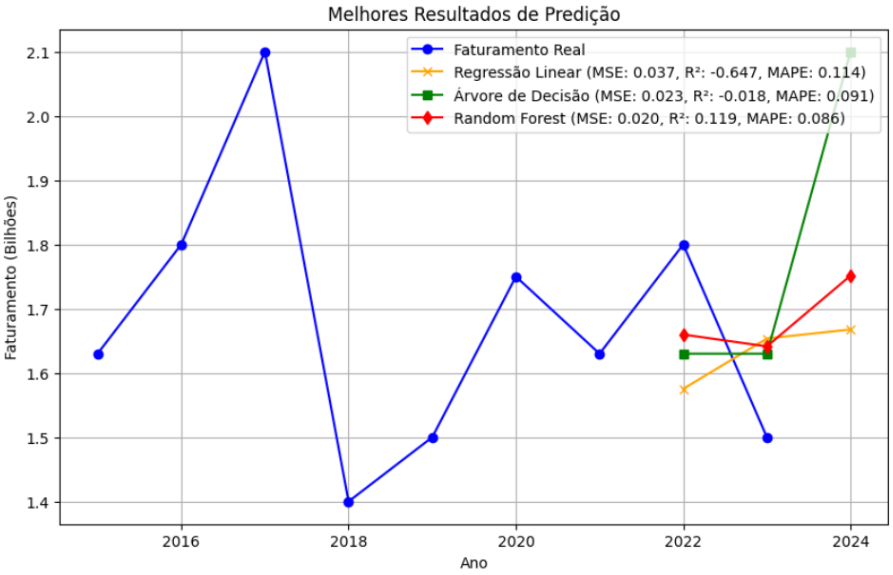
Tabela 4.5: Resultados de desempenho dos modelos após a aplicação do *Grid Search*.

Modelo	**MSE**	R^2	**MAPE**
Regressão Linear	0.037	-0.647	0.114
Árvore de Decisão	0.023	-0.018	0.091
Random Forest	0.020	0.119	0.086

Os resultados indicam que, embora a Regressão Linear tenha apresentado um desempenho razoável sem ajustes de hiperparâmetros, os modelos de Árvore de Decisão e *Random Forest* melhoraram de forma significativa após a busca por hiperparâmetros, especialmente o *Random Forest*, que alcançou o melhor desempenho geral, com MSE de 0.020, R^2 positivo de 0.119 e MAPE de 0.086.

Na Figura 4.5, é possível observar a comparação entre o faturamento real (em azul) e as previsões dos três modelos (anos de teste e predição). O *Random Forest*, que apresentou os melhores resultados, conseguiu prever de maneira mais precisa as tendências de crescimento em 2024, enquanto a Árvore de Decisão e a Regressão Linear tiveram dificuldades em capturar a forte variação de faturamento.

Figura 4.5: Comparação entre o faturamento real e as previsões geradas pelos modelos após a aplicação do Grid Search.



Este experimento reforçou a importância da busca por hiperparâmetros, especialmente para modelos como *Random Forest* e Árvore de Decisão, que se beneficiam de ajustes finos. No próximo experimento, será implementado o método de predições sucessivas, em que os modelos serão reajustados após cada predição, utilizando os resultados reais para simular um re-treinamento anual e melhorar a capacidade dos modelos de capturar padrões temporais.

4.3.4 Experimento 4: Predições Sucessivas

No quarto experimento, foi implementada a técnica de predições sucessivas, onde os modelos são re-treinados após cada predição, utilizando os valores reais de faturamento como parte do re-treinamento. Essa abordagem tem como objetivo simular um cenário em que, ao final de cada ano, os modelos são atualizados com as informações mais recentes, buscando melhorar a capacidade de predição para os anos seguintes.

A ideia por trás dessa abordagem é que, ao incorporar os dados reais a cada ano, os modelos possam capturar melhor as dinâmicas e variações temporais que podem não estar presentes nos dados anteriores. Isso é particularmente útil para cenários onde as tendências podem mudar de forma significativa com o passar do tempo.

Novamente, os três modelos utilizados nos experimentos anteriores foram empregados: Regressão Linear, Árvore de Decisão e *Random Forest*, todos com seus melhores hiperparâmetros, conforme identificado no experimento anterior. Os dados de treinamento

cobriram o período de 2015 a 2021, com as previsões sendo feitas para os anos de 2022, 2023 e 2024.

A técnica de normalização utilizada foi o *Normalizer* para a Regressão Linear e o *MinMaxScaler* para a Árvore de Decisão e o *Random Forest*, conforme identificado nos experimentos anteriores como as técnicas de normalização que apresentaram os melhores resultados para cada modelo.

Os resultados para os anos de 2022 e 2023, medidos com base nas métricas de Erro Quadrático Médio (MSE), Coeficiente de Determinação (R^2) e Erro Percentual Médio Absoluto (MAPE), são apresentados na Table 4.6:

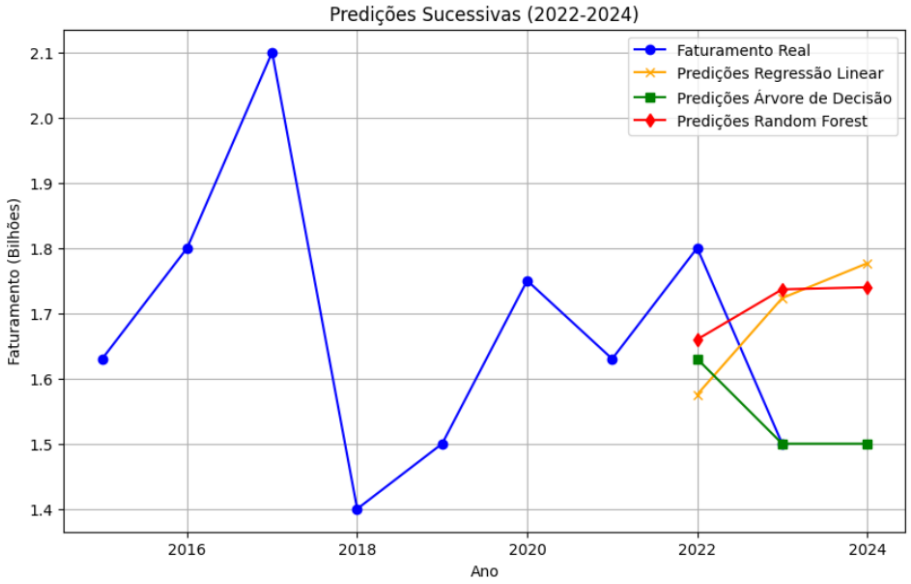
Tabela 4.6: Resultados de Desempenho com Predições Sucessivas (2022-2023).

Modelo	**MSE**	R^2	**MAPE**
Regressão Linear	0.050	-1.229	0.137
Árvore de Decisão	0.014	0.358	0.047
Random Forest	0.038	-0.681	0.118

Conforme observado, a Árvore de Decisão apresentou o melhor desempenho, com um MSE de 0.014, um R^2 positivo de 0.358 e um MAPE de 0.047, o que indica uma capacidade razoável de capturar as variações do faturamento real. O *Random Forest*, embora tenha apresentado um R^2 negativo, teve um MAPE mais baixo (0.118) e uma performance estável. A Regressão Linear, por outro lado, teve o pior desempenho entre os três, com um MSE de 0.050 e um R^2 de -1.229, evidenciando dificuldades em capturar as variações dos anos subsequentes.

Na Figura 4.6, é possível observar a comparação entre as previsões sucessivas para os anos de 2022 a 2024 e o faturamento real. A Árvore de Decisão se mostrou mais próxima da realidade, enquanto a Regressão Linear apresentou maiores desvios.

Figura 4.6: Comparação entre o faturamento real e as previsões geradas pelos modelos de Regressão Linear, Árvore de Decisão e *Random Forest* após predições sucessivas.



Com base nos resultados, conclui-se que este experimento mostrou que a técnica de predições sucessivas pode ser benéfica, especialmente para modelos como a Árvore de Decisão, que conseguiu capturar bem as variações temporais do faturamento da Riot Games. O *Random Forest* também se beneficiou da abordagem, embora os ganhos tenham sido mais discretos. Já a Regressão Linear continuou a apresentar dificuldades em capturar padrões complexos, mesmo com a atualização dos dados. Em resumo, o experimento trouxe ganhos, especialmente para a Árvore de Decisão e o *Random Forest*, em relação aos experimentos anteriores. Esses resultados destacam a importância de utilizar técnicas que permitem a incorporação de dados mais recentes ao longo do tempo, ajustando os modelos conforme novas informações são disponibilizadas.

No próximo experimento, a técnica de redução de dimensionalidade utilizando a *Principal Component Analysis* (PCA) foi introduzida com o objetivo de verificar se a redução no número de variáveis independentes pode melhorar ainda mais a performance dos modelos, ao focar em componentes principais que concentrem a maior parte da variabilidade dos dados.

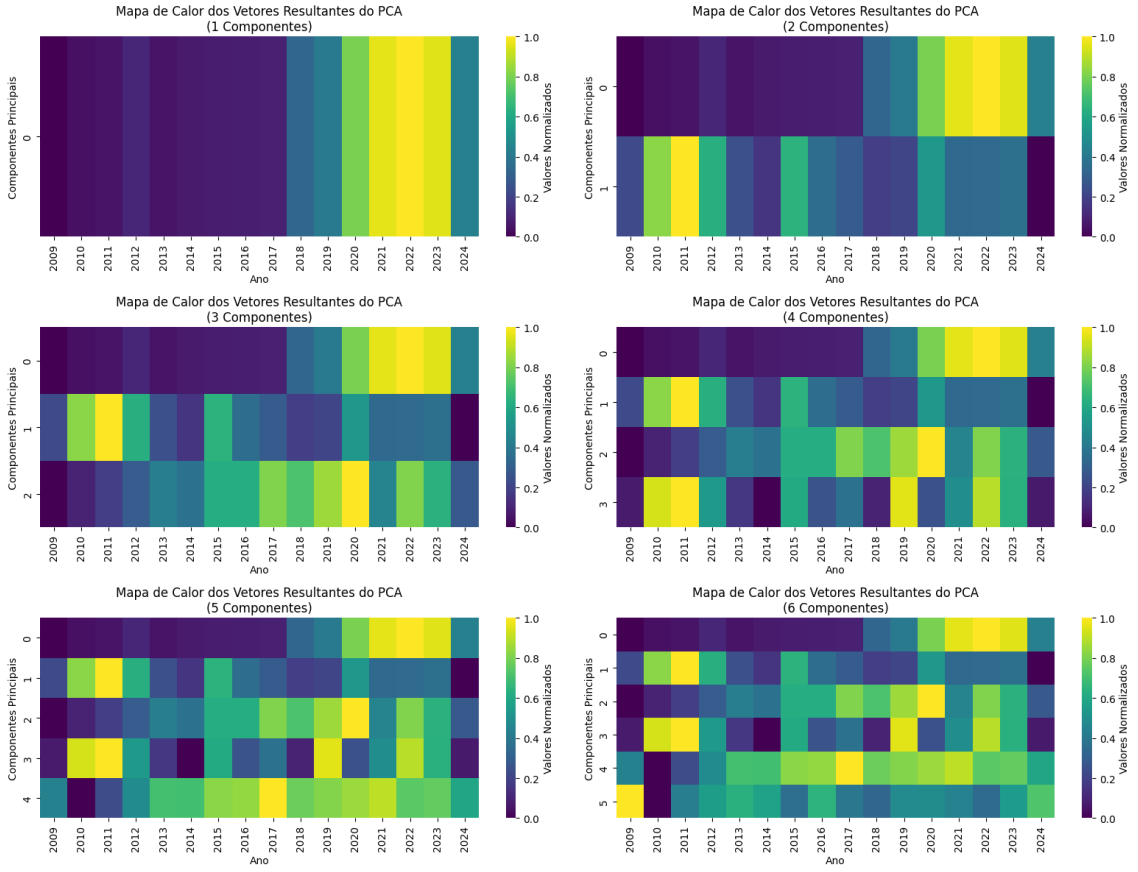
4.3.5 Experimento 5: Predições Sucessivas com PCA

Neste experimento, avaliamos o impacto da redução de dimensionalidade usando a técnica PCA nos modelos de Regressão Linear, Árvore de Decisão e *Random Forest*. O número de componentes principais variou de 1 a 6 para determinar o ajuste ideal. A

razão para isso é que o PCA realiza uma decomposição linear dos dados em componentes ortogonais, e o número máximo de componentes que o PCA pode extrair é igual ao número de amostras, já que a dimensionalidade dos dados não pode ser maior do que o número de amostras disponíveis para calcular as variâncias (Alpaydin, 2020).

Nos vetores originais, representados na Figura 4.1, cada variável possui um significado específico, como descrito na Seção 4.2. Essas variáveis possuem escalas diferentes, e a normalização foi aplicada para garantir uma comparabilidade justa entre elas. O gráfico revela padrões específicos de comportamento ao longo dos anos para cada variável individual. Por outro lado, a Figura 4.7 representa os resultados da aplicação da PCA. Observe que os dados foram projetados em um novo espaço dimensional, no qual as variáveis originais foram combinadas para formar componentes ortogonais. Nesses gráficos, cada componente principal não possui um significado específico relacionado diretamente às variáveis originais, mas representa uma combinação das variáveis que conserva o máximo de variância possível dos dados. A aplicação do PCA permite reduzir a dimensionalidade dos dados, eliminando a multicolinearidade e destacando as relações mais importantes entre as variáveis (Müller; Guido, 2016).

Figura 4.7: Representação dos vetores resultantes em mapa de calor, por ano, após a aplicação do PCA.



Seguindo com o experimento, os modelos foram treinados com os vetores resultantes, sendo realizadas predições sucessivas para os anos de 2022 a 2024. Os resultados são apresentados na Tabela 4.7.

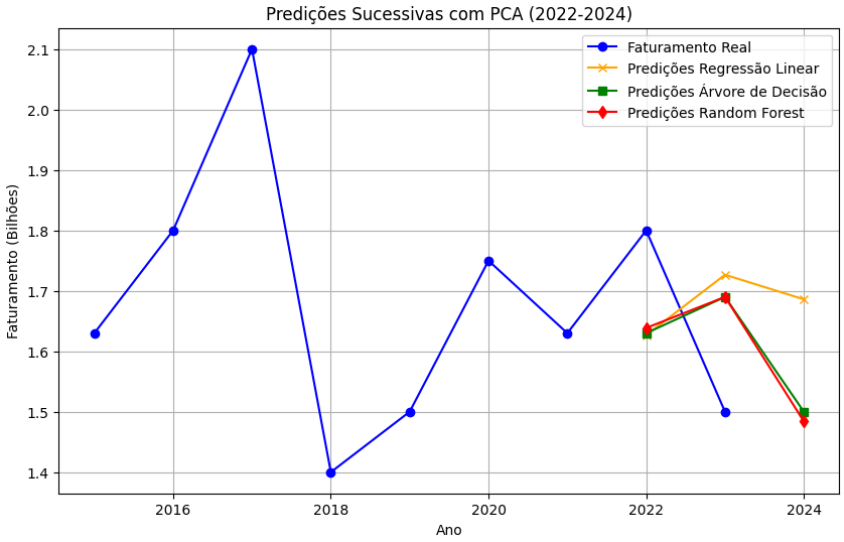
Tabela 4.7: Resultados das métricas para diferentes números de componentes PCA.

Método	**Métrica**	**1 Comp.**	**2 Comp.**	**3 Comp.**	**4 Comp.**	**5 Comp.**	**6 Comp.**
Regressão Linear	MSE	0.041	0.027	0.025	0.027	0.041	0.028
	R^2	-0.809	-0.218	-0.128	-0.185	-0.814	-0.251
	MAPE	0.124	0.098	0.098	0.100	0.122	0.074
Árvore de Decisão	MSE	0.033	0.017	0.029	0.029	0.053	0.023
	R^2	-0.444	0.238	-0.284	-0.284	-1.376	-0.018
	MAPE	0.111	0.077	0.104	0.104	0.127	0.091
Random Forest	MSE	0.031	0.031	0.027	0.022	0.040	0.029
	R^2	-0.381	-0.372	-0.200	0.004	-0.759	-0.284
	MAPE	0.108	0.107	0.099	0.092	0.118	0.094

Com base nesses resultados, segue uma análise geral do desempenho de cada modelo combinado com um número específico de componentes.

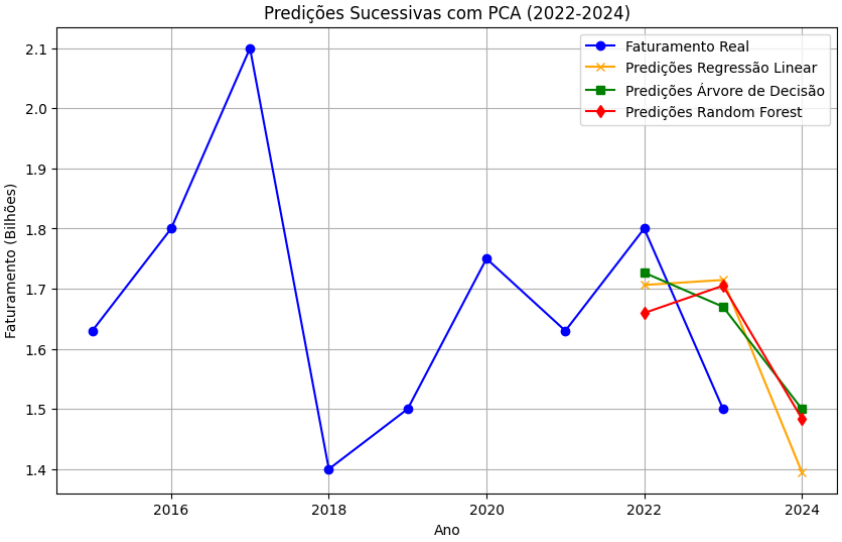
- **Vetores com 1 componente**: *Random Forest* apresentou o melhor desempenho com menor MSE (0.031), menor MAPE (10,8%) e R^2 menos negativo (-0.381), sugerindo que é o modelo mais adequado para capturar as variações do faturamento. A Árvore de Decisão teve um desempenho intermediário (MSE de 0.033 e MAPE de 11,1%), enquanto a Regressão Linear apresentou o pior desempenho, com maior MSE (0.041) e um R^2 muito negativo (-0.809), o que reflete as predições mais distantes dos valores reais. No gráfico apresentado na Figura 4.8, observa-se que as predições da *Random Forest* e da Árvore de Decisão são mais próximas do valor real, apesar de não fornecerem resultados suficientes.

Figura 4.8: Predições Sucessivas com PCA (1 componente)



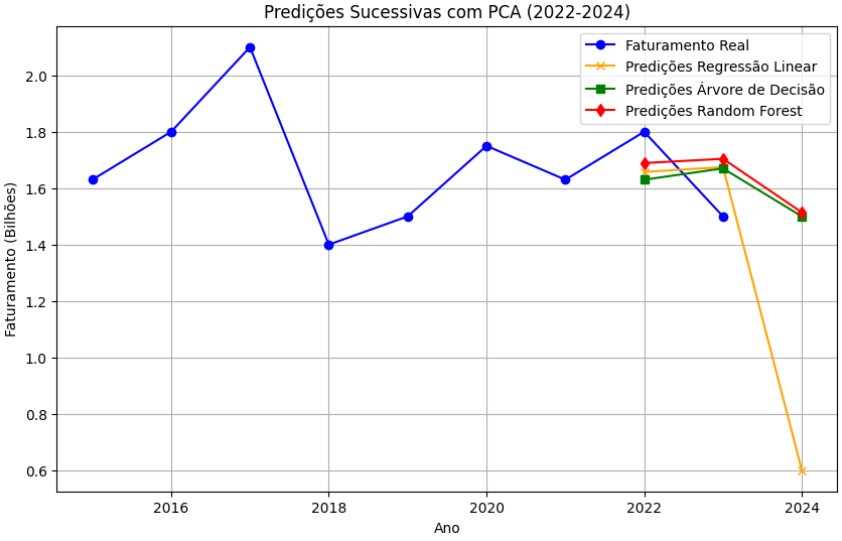
- **Vetores com 2 componentes**: Com duas componentes PCA, a Árvore de Decisão apresentou o melhor desempenho com MSE de 0.017 e R^2 positivo (0.238), mostrando um bom ajuste. A *Random Forest* teve um desempenho moderado (MSE de 0.031), mas com R^2 ainda negativo (-0.372). A Regressão Linear melhorou levemente, mas ainda apresenta R^2 negativo (-0.218) e subestima os valores, especialmente para 2024, que é uma previsão (Figura 4.9).

Figura 4.9: Predições Sucessivas com PCA (2 componentes)



- **Vetores com 3 componentes**: Com 3 componentes os três métodos apresentaram resultados semelhantes para os anos de 2022 e 2023, enquanto a Regressão Linear perde a estabilidade e prevê uma queda acentuada no faturamento, inconsistente com a realidade. Assim, Árvore de Decisão e *Random Forest* permanecem as opções mais adequadas para esse cenário. Seguem as previsões na Figura 4.10.

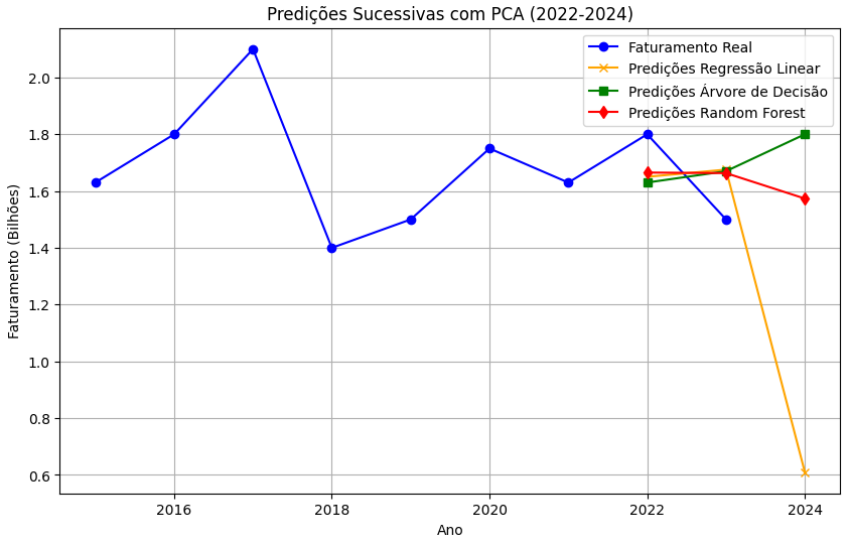
Figura 4.10: Predições Sucessivas com PCA (3 componentes)



- **Vetores com 4 componentes**: Árvore de Decisão apresenta um MAPE de 10,4%, enquanto a *Random Forest* tem um MAPE de 9,2%, ambos com predições próximas

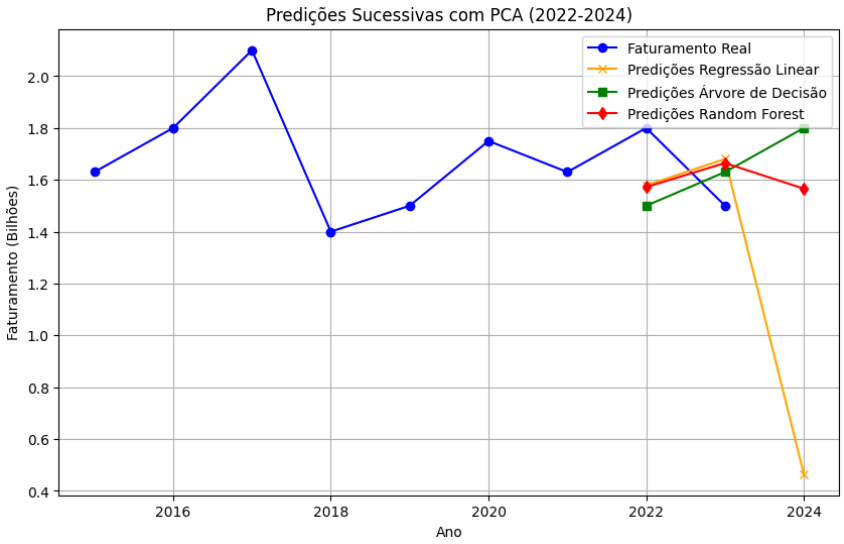
dos valores reais em 2022 e 2023, e uma leve divergência em 2024. A Regressão Linear, por outro lado, mostra um MAPE de 10%, mas com um comportamento de subestimação severa, com uma queda acentuada da previsão em 2024, conforme ilustrado pela linha amarela no gráfico. Em resumo, Árvore de Decisão e *Random Forest* continuam sendo as melhores opções, enquanto a Regressão Linear falha em capturar corretamente a tendência para o ano de 2024, conforme Figura 4.11.

Figura 4.11: Predições Sucessivas com PCA (4 componentes)



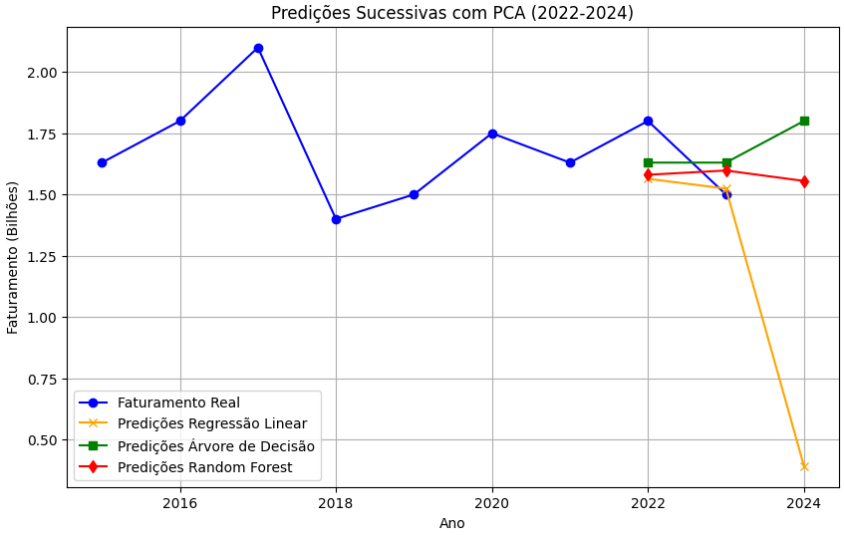
- **Vetores com 5 componentes**: A figura 4.12 mostra que a performance piorou significativamente com 5 componentes. A Regressão Linear continua prevendo uma queda acentuada para 2024, com MSE de 0.041 e R^2 de -0.814, enquanto a Árvore de Decisão atingiu seu pior resultado, com MSE de 0.053 e R^2 de -1.376, o que pode indicar *overfitting* do modelo.

Figura 4.12: Predições Sucessivas com PCA (5 componentes)



- **Vetores com 6 componentes**: A adição de 6 componentes trouxe pouca ou nenhuma melhoria significativa em relação à análise anterior. O *Random Forest* manteve uma performance aceitável (MSE de 0.029), mas os ganhos foram limitados. A Regressão Linear teve um desempenho mediano com MSE de 0.028 e MAPE de 0.074, continuando com uma queda acentuada na previsão para 2024 extremamente inconsistente.

Figura 4.13: Predições Sucessivas com PCA (6 componentes)



Fica assim demonstrado que a redução de dimensionalidade usando a técnica PCA

teve um impacto significativo no desempenho dos modelos, especialmente para os mais complexos, como *Random Forest* e Árvore de Decisão. O *Random Forest*, com 4 componentes principais (Figura 4.11), foi o modelo que melhor capturou as tendências de crescimento do faturamento, apresentando o melhor desempenho geral em termos de MSE, R^2 e MAPE. Já a Árvore de Decisão alcançou seu melhor resultado com 2 componentes (Figura 4.9), mostrando uma boa capacidade de predição com menos variáveis.

O PCA se mostrou eficaz para melhorar a capacidade dos modelos de lidar com as variações nos dados, especialmente nos modelos que se beneficiam da combinação de múltiplas variáveis. No entanto, a escolha do número de componentes foi crucial, visto que a Regressão Linear apresentou desempenho insatisfatório quando o número de componentes era inadequado, seja por ser excessivo ou insuficiente.

Além disso, um ponto importante é a interpretação visual e predição de 2024. Embora as métricas da Árvore de Decisão com 2 componentes tenham sido superiores até 2023, o *Random Forest* com 4 componentes apresentou uma previsão mais alinhada à tendência de crescimento do faturamento para 2024, conforme observado na comparação das Figuras 4.9 e 4.11. Isso demonstra que, apesar das diferenças numéricas nas métricas até 2023, o *Random Forest* ofereceu uma previsão mais realista para o futuro, o que é um fator determinante para escolher o modelo mais adequado.

4.4 Discussão dos Resultados

Os resultados apresentados fornecem *insights* importantes sobre a eficácia dos diferentes modelos preditivos para estimar o faturamento da Riot Games, com base em dados históricos de lançamentos e atualizações de campeões e *skins* de League of Legends. Cada experimento revelou características valiosas sobre o comportamento dos modelos diante de variações de pré-processamento, normalização e redução de dimensionalidade.

O desempenho superior do modelo *Random Forest*, especialmente após a normalização e a busca por hiperparâmetros, destaca sua robustez em capturar padrões complexos nos dados, mesmo em cenários sem normalização ou com menor quantidade de componentes principais via PCA. Esse resultado está em linha com a literatura, que aponta o *Random Forest* como um método capaz de lidar com variáveis de diferentes escalas e naturezas, tornando-o adequado para cenários de alta dimensionalidade e dados heterogêneos, como no caso de variáveis envolvendo tanto eventos quanto lançamentos de *skins* e campeões.

A Árvore de Decisão, embora tenha apresentado bons resultados em alguns cenários, mostrou-se menos consistente, com sua performance flutuando dependendo da quantidade de componentes principais usadas. Isso sugere que, enquanto esse modelo consegue capturar bem padrões específicos, ele é mais suscetível a *overfitting* e precisa de ajustes finos para entregar resultados mais estáveis. No entanto, seu desempenho no experimento de predi-

ções sucessivas foi promissor, indicando que, quando atualizado com dados mais recentes, pode fornecer boas estimativas para previsões a curto prazo.

Já a Regressão Linear teve dificuldades em capturar as dinâmicas não lineares dos dados, conforme evidenciado por seus baixos valores de R^2 e maiores erros. Mesmo com normalização e técnicas de redução de dimensionalidade, o modelo não conseguiu melhorar substancialmente, o que indica sua limitação para esse tipo de problema.

Além disso, a aplicação do PCA foi útil para reduzir a dimensionalidade dos dados sem perda significativa de informações. No entanto, o número de componentes principais teve um impacto considerável no desempenho dos modelos, sendo o *Random Forest* com quatro componentes o melhor cenário encontrado. Essa redução permitiu que o modelo se concentrasse nos fatores mais relevantes, minimizando ruídos e facilitando a interpretação dos resultados.

Por fim, um ponto notável é a tendência decrescente observada nos lançamentos de novos campeões ao longo dos anos, especialmente a partir de 2015. Esse padrão reflete uma possível mudança na estratégia da Riot Games, com maior foco em atualizações e lançamentos de *skins*, o que pode influenciar diretamente as estimativas de faturamento e as predições futuras.

Os achados desta pesquisa indicam que o uso de técnicas de aprendizado de máquina, como o *Random Forest* combinado com a redução de dimensionalidade, oferece um potencial robusto para prever o faturamento da Riot Games com base nos dados históricos. No entanto, os resultados também sugerem que, devido à complexidade do fenômeno modelado e à quantidade limitada de dados disponíveis, as predições apresentam uma margem de erro significativa. Embora o *Random Forest* tenha capturado bem as tendências passadas, ajustes periódicos, como nas predições sucessivas, são essenciais para melhorar a capacidade de capturar variações no comportamento do mercado e nas estratégias de lançamento da empresa, garantindo maior precisão nas previsões para períodos futuros.

5 Conclusão

O presente trabalho teve como objetivo principal desenvolver um modelo preditivo capaz de estimar o faturamento anual da Riot Games com League of Legends, utilizando técnicas de *Machine Learning* e dados históricos relacionados aos conteúdos lançados no jogo. Para atingir esse propósito, foram estabelecidos objetivos específicos que incluíram a coleta e organização de dados públicos sobre campeões, *skins* e eventos, o pré-processamento desses dados, a implementação e comparação de diferentes modelos preditivos, a aplicação de técnicas de redução de dimensionalidade e a análise dos resultados obtidos.

Ao longo do estudo, foi possível alcançar de forma satisfatória os objetivos propostos. A coleta de dados, embora desafiadora devido à indisponibilidade de informações financeiras oficiais da Riot Games, foi realizada com sucesso a partir de fontes públicas como o *League of Legends Fandom* e estimativas de faturamento disponíveis em sites especializados e fóruns da comunidade. O pré-processamento dos dados permitiu a construção de vetores de características adequados para os modelos de aprendizado de máquina, contemplando variáveis relevantes como número de campeões e *skins* lançados, características dos campeões e eventos associados.

A implementação dos modelos de Regressão Linear, Árvore de Decisão e *Random Forest* possibilitou a comparação de diferentes abordagens preditivas. Os resultados indicaram que o modelo *Random Forest*, especialmente quando combinado com técnicas de normalização e redução de dimensionalidade via PCA, apresentou o melhor desempenho na previsão da tendência do faturamento anual, com erro percentual de 7.49% para 2022 e 10.88% para 2023. Apesar das limitações inerentes aos dados utilizados, os modelos conseguiram captar tendências gerais, oferecendo *insights* valiosos sobre os fatores que podem influenciar a receita da Riot Games.

Entretanto, é importante reconhecer as limitações deste estudo. Os dados de faturamento utilizados são estimativas obtidas em sites de notícias e fóruns, o que pode resultar em discrepâncias significativas em relação à realidade financeira da empresa. Essa imprecisão nos dados de saída impacta diretamente a precisão dos modelos preditivos, limitando a confiabilidade das conclusões. Além disso, a alta dimensionalidade dos vetores de variáveis independentes, com 179 características, pode ter introduzido ruídos e complexidades adicionais na modelagem.

Apesar dessas limitações, o estudo possui grande importância do ponto de vista didático e metodológico. Ele demonstra a aplicabilidade de técnicas de *Machine Learning* em contextos com dados limitados ou parcialmente indisponíveis, abrindo margem para reflexões e pesquisas futuras. A seguir, são sugeridas algumas direções para trabalhos

futuros:

- **Agrupamento de Eventos**: Uma possibilidade é agrupar mais eventos e temas de *skins*, reduzindo a dimensionalidade dos vetores de variáveis independentes. Essa abordagem pode simplificar o modelo, facilitar a interpretação dos resultados e diminuir o risco de *overfitting*.

- **Incorporação de Novas Estatísticas**: Integrar outras estatísticas disponíveis, como número de jogadores ativos, partidas jogadas, taxas de vitória e engajamento dos usuários. Dados provenientes de plataformas como o OP.GG ou da API[1] oficial da Riot Games podem enriquecer o conjunto de variáveis e potencialmente aumentar a precisão dos modelos preditivos.

- **Consideração dos Custos em RP por *Skin***: Reavaliar a inclusão dos custos em Riot Points (RP) de cada *skin* no modelo. Embora tenham sido inicialmente excluídos para evitar desbalanceamento, uma abordagem que normalize ou categorize esses valores pode fornecer informações adicionais sobre a contribuição de *skins* de diferentes valores para o faturamento.

- **Estudos de Correlação e Variância**: Realizar análises estatísticas aprofundadas, como estudos de correlação e análise de variância, para identificar quais variáveis possuem maior influência sobre o faturamento. Compreender a correlação entre as variáveis independentes e a variável dependente pode orientar a seleção de características mais relevantes e aprimorar os modelos preditivos.

A partir dos resultados obtidos, conclui-se que, embora o problema de pesquisa tenha sido abordado e os objetivos alcançados em termos metodológicos, a precisão das previsões é limitada pela qualidade e disponibilidade dos dados. A falta de dados financeiros oficiais da Riot Games impede a validação robusta dos modelos e a confirmação das tendências identificadas. No entanto, o estudo contribui significativamente para demonstrar a viabilidade de aplicar técnicas de *Machine Learning* na previsão de faturamento em contextos similares.

Como considerações finais, ressalta-se a importância de continuar explorando métodos que permitam estimar o desempenho financeiro de empresas em setores com informações restritas. A indústria de jogos eletrônicos, em constante crescimento e evolução, oferece um campo fértil para pesquisas que integrem análise de dados, aprendizado de máquina e compreensão de mercados digitais. Trabalhos futuros que superem as limitações

[1]Uma *Application Programming Interface* (API) é um conjunto de ferramentas, definições e protocolos que permitem que aplicativos e softwares diferentes se conectem e interajam (Pereira, 2016).

aqui identificadas poderão não apenas aprimorar as previsões financeiras, mas também fornecer *insights* estratégicos valiosos para empresas e pesquisadores.

Referências

ALEXANDER, L. **Don't Monetize Like I, League of Legends, I Consultant Says**. 2023. Website. Acesso em 09 de setembro de 2024. Disponível em: <https://www.gamedeveloper.com/business/don-t-monetize-like-i-league-of-legends-i-consultant-says>.

ALPAYDIN, E. **Introduction to Machine Learning, fourth edition**. [S.l.]: MIT Press, 2020. (Adaptive Computation and Machine Learning series). ISBN 9780262043793.

ALVES, P. C. P.; PRADO, S. C. C. Estudo comparativo entre algoritmos de machine learning aplicados à previsão de séries temporais do mercado financeiro. In: **Congresso de Tecnologia-Fatec Mococa**. [S.l.: s.n.], 2022. v. 6, n. 2.

ARIK, K. Machine learning models on moba gaming: league of legends winner prediction. **Acta Infologica**, Istanbul University, v. 7, n. 1, p. 139–151, 2023.

BRAZDIL, P. Construção de modelos de decisão a partir de dados. 1999.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001.

BROWN, S. **Machine Learning Explained**. 2021. Acesso em: 03 de setembro de 2024. Disponível em: <https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained>.

BRYNJOLFSSON, E.; MITCHELL, T.; ROCK, D. What can machines learn and what does it mean for occupations and the economy? In: AMERICAN ECONOMIC ASSOCIATION 2014 BROADWAY, SUITE 305, NASHVILLE, TN 37203. **AEA papers and proceedings**. [S.l.], 2018. v. 108, p. 43–47.

CORMEN, T. **Desmistificando algoritmos**. [S.l.]: Elsevier Editora Ltda., 2017. ISBN 9788535271799.

CORRÊA, J. **Riot Games: Uma história de sucesso baseada em ouvir a comunidade**. 2020. Acesso em 02 de julho de 2024. Disponível em: <https://ge.globo.com/e-sportv/noticia/riot-games-uma-historia-de-sucesso-baseada-em-ouvir-a-comunidade.ghtml>.

DACANAY, R. **How Much Money Did League of Legends Make in 2021?** 2021. DBLTAP, site de notícias. Acesso em: 12 de setembro de 2024. Disponível em: <https://www.dbltap.com/posts/how-much-money-did-league-of-legends-make-in-2021-01fr6hfexgdt>.

FANDOM, L. of L. **Module: SkinData**. 2024. Acesso em: setembro de 2024. Disponível em: <https://leagueoflegends.fandom.com/wiki/Module:SkinData/data?action=edit>.

FORTI, M. **Técnicas de machine learning aplicadas na recuperação de crédito do mercado brasileiro**. 2018. Tese (Doutorado), 2018.

GARCIA, S. C. O uso de árvores de decisão na descoberta de conhecimento na área da saúde. 2003.

GILL, S. **League of Legends Data**. 2023. Website. Acesso em 09 de setembro de 2024. Disponível em: <https://prioridata.com/data/league-of-legends/>.

GOMES, L. F. A. M.; GOMES, C. F. S. **Tomada de decisão gerencial: enfoque multicritério .** [S.l.]: Editora Atlas SA, 2000.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>.

HASLWANTER, T. **An Introduction to Statistics with Python: With Applications in the Life Sciences**. [S.l.]: Springer International Publishing, 2022. (Statistics and Computing). ISBN 9783030973711.

HYNDMAN, R. **Forecasting: principles and practice**. [S.l.]: OTexts, 2018.

JAMES, G. **An introduction to statistical learning**. [S.l.]: springer, 2013.

MALONE, T. W.; RUS, D.; LAUBACHER, R. Artificial intelligence and the future of work. **A report prepared by MIT Task Force on the work of the future, Research Brief**, v. 17, p. 1–39, 2020.

MEDIAS, F. L. **Esports-League of Legends revenue to hit $1 billion this year, data suggests**. 2021. Disponível em: <https://web.archive.org/web/20210117093230/https://www.reuters.com/article/esports-lol-revenue-idUSFLM2vzDZL>.

Mobile Marketing Reads. **League of Legends Revenue and User Stats**. 2024. Acessado em: 16 de setembro de 2024. Disponível em: <https://mobilemarketingreads.com/league-of-legends-revenue-and-user-stats/>.

MOHRI, M. **Foundations of machine learning**. [S.l.]: MIT press, 2018.

MÜLLER, A.; GUIDO, S. **Introduction to Machine Learning with Python: A Guide for Data Scientists**. [S.l.]: O'Reilly Media, Incorporated, 2016. ISBN 9781449369415.

OTHMANE, A.; YOUKANA, I.; KAHLOUL, L.; BOUREKKACHE, S. Ai-based prediction for glucose levels: A comparative study of machine learning and deep learning approaches. In: **Proceedings of the International Conference on Emerging Intelligent Systems for Sustainable Development (ICEIS 2024)**. [S.l.]: Atlantis Press, 2024. p. 231–245. ISBN 978-94-6463-496-9. ISSN 1951-6851.

PAULA, M. B. d. et al. Indução automática de árvores de decisão. Florianópolis, SC, 2002.

PEDREGOSA, F. e. a. **Scikit-learn: Machine Learning em Python - Avaliação de Modelos**. 2011. Acesso em: 2024-09-19. Disponível em: <https://scikit-learn.org/stable/modules/model_evaluation.html#the-scoring-parameter-defining-model-evaluation-rules>.

PEREIRA, C. **Construindo APIs REST com Node.js: Caio Ribeiro Pereira**. [S.l.]: Casa do Código, 2016. ISBN 9788555191510.

Public Company Accounting Oversight Board (PCAOB). **AT Section 301: Financial Forecasts and Projections**. 2001. <https://pcaobus.org/oversight/standards/attestation-standards/details/AT301>. Acesso em: 03 de setembro de 20241.

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence: a modern approach**. [S.l.]: Pearson, 2016.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of research and development**, IBM, v. 3, n. 3, p. 210–229, 1959.

SINGH, D.; SINGH, B. Investigating the impact of data normalization on classification performance. **Applied Soft Computing**, v. 97, p. 105524, 2020. ISSN 1568-4946.

SMIT, R. A machine learning approach for recommending items in league of legends. **Netherlands: Radboud University**, 2019.

SUPERDATA. **Market Brief Year in Rreview**. 2018. Acesso em: 03 de setembro de 2024. Disponível em: <https://web.archive.org/web/20190121010757/https://www.superdataresearch.com/market-data/market-brief-year-in-review/>.

UNPINGCO, J. **Python for Probability, Statistics, and Machine Learning**. [S.l.]: Springer International Publishing, 2022. ISBN 9783031046483.

VASCONCELOS, S. Análise de componentes principais (pca). **Rio de Janeiro**, 2007.

ZIPPIA. **Riot Games - Receita, Zippia**. 2024. Acessado em: 18 de setembro de 2024. Disponível em: <https://www.zippia.com/riot-games-careers-36774/revenue/>.