



inteli

GIA  
Rede Gazeta



## Controle do Documento

### Histórico de revisões

| Data       | Autor                                      | Versão | Resumo da atividade  |
|------------|--|--------|--|
| 09/08/2022 | João Marques e Raphael Antunes             | 0.1    | Canvas de proposta de valor  |
| 04/08/2022 | Raduan Muarrek e Luana Parra               | 0.1.2  | Análise de indústria   |
| 04/08/2022 | Vitor Oliveira e João Suarez               | 0.1.3  | Matriz de riscos   |
| 08/08/2022 | João Marques, Raphael Antunes, Luana Parra | 0.1.4  | Matriz SWOT  |
| 09/08/2022 | João Marques e Raphael Antunes             | 0.1.5  | Canvas de proposta de valor  |
| 10/08/2022 | Raduan Muarrek e João Marques              | 0.1.6  | Descrição da predição  |
| 10/08/2022 | Luana Parra e Vitor Oliveira               | 0.1.7  | Descrição dos dados  |
| 11/08/2022 | Elisa Flemer e Raphael Antunes             | 0.1.8  | Criação dos gráficos para a análise de dados                       |
| 11/08/2022 | Luana Parra                                | 0.1.9  | Revisão dos artefatos da Sprint 1                                  |
| 12/08/2022 | Elisa Flemer                               | 0.2.0  | Revisão final para Sprint 1  |
| 16/08/2022 | João Suarez e Elisa Flemer                 | 0.2.1  | Correção das legendas dos gráficos e início da preparação de dados |
| 28/08/2022 | Luana Parra e Elisa Flemer                 | 1.0    | Desenvolvimento e revisão dos artefatos da Sprint 2.               |

# Sumário

|  |           |
|--|-----------|
| <b>1. Introdução</b>   | <b>5</b>  |
| <b>2. Objetivos e Justificativa</b>                                | <b>6</b>  |
| 2.1. Objetivos   | 6         |
| 2.2. Justificativa   | 6         |
| <b>3. Metodologia</b>  | <b>7</b>  |
| 3.1. CRISP-DM  | 7         |
| 3.2. Ferramentas   | 7         |
| 3.3. Principais técnicas empregadas                                | 8         |
| <b>4. Desenvolvimento e Resultados</b>                             | <b>8</b>  |
| 4.1. Compreensão do Problema                                       | 8         |
| 4.1.1. Contexto da indústria                                       | 8         |
| <b>Análise aprofundada seguindo o modelo de 5 forças de Porter</b> | <b>9</b>  |
| <b>4.1.2. Análise SWOT</b>   | <b>11</b> |
| 4.1.3. Planejamento Geral da Solução                               | 11        |
| 4.1.4. Value Proposition Canvas                                    | 12        |
| 4.1.5. Matriz de Riscos  | 13        |
| <b>4.1.6. Personas</b>   | <b>14</b> |
| <b>4.1.7. Jornadas do Usuário</b>                                  | <b>16</b> |
| <b>4.2. Compreensão dos Dados</b>                                  | <b>17</b> |
| <b>Análise preliminar dos dados</b>                                | <b>18</b> |
| <b>Audiência por hora por emissora (Seg a Sex)</b>                 | <b>18</b> |
| <b>Audiência por hora por emissora (Sábado)</b>                    | <b>19</b> |
| <b>Audiência por hora por emissora (Domingo)</b>                   | <b>20</b> |
| <b>Audiência da Emissora 0 por dia da semana</b>                   | <b>21</b> |
| <b>Audiência por dia do mês por emissora</b>                       | <b>24</b> |
| <b>Audiência da Emissora 0 por mês</b>                             | <b>25</b> |
| <b>Audiência por emissora por ano</b>                              | <b>26</b> |

|   |           |
|---|-----------|
| Audiência da Emissora 0 por classe socioeconômica | 27        |
| Audiência da Emissora 0 por faixa etária          | 28        |
| Audiência total da Emissora 0 por gênero          | 28        |
| Audiência da Emissora 0 por gênero                | 30        |
| Considerações sobre o resultado desejado          | 30        |
| <b>4.3. Preparação dos Dados</b>                  | <b>31</b> |
| Pré-processamento dos dados                       | 31        |
| Anonimização dos dados                            | 31        |
| Otimização de processamento dos arquivos          | 32        |
| Formatação de datas                               | 32        |
| Merge com grade horária                           | 33        |
| Checando a existência de outliers                 | 33        |
| Checando valores nulos e ausentes                 | 34        |
| Checando categorias de baixa frequência           | 34        |
| Seleção de features                               | 35        |
| Teste de hipóteses                                | 36        |
| Hipótese 1  | 37        |
| Hipótese 2  | 37        |
| Hipótese 3  | 37        |
| 4.4. Modelagem                                    | 38        |
| 4.5. Avaliação                                    | 39        |
| <b>5. Conclusões e Recomendações</b>              | <b>41</b> |
| <b>6. Referências</b>                             | <b>42</b> |
| <b>Anexos</b>                                     | <b>43</b> |

# 1. Introdução

A Rede Gazeta de Comunicações é a maior empresa de comunicação do Espírito Santo, com mais de 500 funcionários. Fundada em 1928 com o jornal *A Gazeta*, tem como maior propósito informar, entreter e prestar serviços de comunicação aos capixabas com qualidade, ética e inovação, contribuindo para o desenvolvimento socioeconômico, cultural e de cidadania. Atualmente, o grupo é formado por um site de notícias ([www.agazeta.com.br](http://www.agazeta.com.br)); oito rádios; quatro emissoras de TV aberta (TV Gazeta) afiliadas à Rede Globo; e dois portais de notícias locais (G1 Espírito Santo e o Globo Esporte Espírito Santo).

No que tange às emissoras, devido à sua conexão com a Rede Globo, devem coordenar a programação local com os eventos nacionais produzidos pela rede. Assim, com slots limitados para veicular seus próprios conteúdos, torna-se imprescindível prever a audiência de novos eventos com acurácia para maximizar o retorno de seus investimentos. Objetiva-se, com isso, escolher sempre o melhor evento para cada horário disponível e alocar recursos eficientemente, priorizando campanhas de marketing para programas com menor audiência esperada.

No momento, apesar de existirem registros históricos de score de audiência para diferentes horários e demográficos, não há análises de correlação entre programa (com suas características principais) e slot. Consequentemente, faltam evidências para prever e justificar a transmissão de certos eventos em certos horários, causando significativa incerteza para os stakeholders da empresa a cada modificação na grade de programação.

## 2. Objetivos e Justificativa

### 2.1. Objetivos

O principal objetivo da Rede Gazeta é maximizar o retorno financeiro de cada evento veiculado. Essa monetização se dá através da venda de slots publicitários, cuja precificação, por sua vez, é fortemente influenciada do score de audiência do programa sendo veiculado durante o slot comercial. Assim, em termos mais específicos, para maximizar seus retornos, a Rede Gazeta deve maximizar a audiência de cada conteúdo a fim de atrair mais anunciantes.

### 2.2. Justificativa

A GIA (Gazeta Inteligência Artificial) é um modelo preditivo inovador que calcula o score de audiência esperado, geral e por demográfico, para novos programas em diferentes dias e horários. É disruptiva na indústria por aplicar machine learning para o grande volume de dados produzido pelo Kantar IBOPE e oferecer uma análise detalhada substanciando a previsão de audiência. Antes dele. Assim, a GIA não só produz um diagnóstico objetivo, baseado em tendências históricas, como também elenca as variáveis utilizadas para se chegar a esse resultado, a fim de que o usuário possa validar a linha de raciocínio.

Essas funcionalidades são desconhecidas na indústria televisiva, que costuma delegar essas tarefas inteiramente para funcionários. Estes acabam dedicando muita energia mental e tempo para a análise e tomada de decisão manuais, sendo, portanto, muito mais vulneráveis a erros de interpretação e viés pessoal.

Assim, a GIA é um algoritmo destinado a alavancar a produtividade, assertividade e acurácia do Departamento de Programação da Gazeta, colocando-a, desse modo, a frente dos concorrentes no que tange à escolha e agendamento de novos eventos. A médio e longo prazo, espera-se que as sugestões da GIA, quando aplicadas, aumentem significativamente o sucesso de audiência da emissora e o retorno financeiro de seus investimentos.

## 3. Metodologia

Descreva as etapas metodológicas que foram utilizadas para o desenvolvimento, citando o referencial teórico. Você deve apenas enunciar os métodos, sem dizer ainda como ele foi aplicado e quais resultados obtidos.

### 3.1. CRISP-DM

Descreva brevemente a metodologia CRISP-DM e suas etapas de processo

O CRISP-DM (Cross Industry Standard Process for Data Mining) é uma metodologia ágil voltada para projetos envolvendo Machine Learning, mineração e análise de dados. Sendo assim, é um processo cíclico dividido nas seguintes etapas:

- 1."Business Understanding"(estudo do projeto ou negócio, atendendo os objetivos e interesses do cliente);
- 2."Data Understanding (identificar, coletar e analisar os conjuntos de dados que podem ajudar a atingir os objetivos do projeto);
- 3."Data Preparation" (ocorre a manipulação de dados, filtrando quais dados serão usados para a modelagem);
- 4."Modeling" (desenvolver um modelo e selecionar a técnica de modelagem);
- 5."Evaluation" (avaliar a qualidade, fidedignidade e segurança dos resultados obtidos da etapa de Modelagem);
- 6."Deployment" (se inicia o processo de desenvolvimento dos modelos criados e avaliados nas etapas anteriores)

### 3.2. Ferramentas

Descreva brevemente as ferramentas utilizadas e seus papéis (Google Collaboratory)

| Nome                 | O que é  | Em que foi utilizado | Versão |
|----------------------|--|----------------------|--------|
| Google Collaboratory | É um serviço de nuvem gratuito hospedado pelo próprio Google para incentivar a pesquisa de Aprendizado de Máquina e Inteligência Artificial. |                      |        |

|  |  |  |  |
|--|--|--|--|
|  |  |  |  |
|--|--|--|--|

### 3.3. Principais técnicas empregadas

Descreva brevemente as principais técnicas empregadas, algoritmos e seus benefícios

## 4. Desenvolvimento e Resultados

### 4.1. Compreensão do Problema

#### 4.1.1. Contexto da indústria

A indústria televisiva aberta opera, primordialmente, em modelo B2B através da venda de slots publicitários para anunciantes diversos nos intervalos de cada evento (publicidades, merchandising, comerciais). A receita dessa atividade é diretamente relacionada ao score de audiência atingido pelos programas veiculados pela emissora. Nesse sentido, a atenção dos telespectadores é a mercadoria oferecida a seus clientes, isto é, às empresas que buscam divulgar seus produtos e serviços para um público de larga escala.

Dentre os principais players do mercado, tem a Globo, RecordTV, SBT, Band e TV Brasil. Destes, a Globo apresenta as maiores taxas de audiência em quase todas as situações, beirando um monopólio da indústria. A lista abaixo apresenta mais detalhes sobre cada uma delas.

**Globo:** Assistida por mais de 200 milhões de pessoas diariamente (no Brasil e no exterior), é a segunda maior rede de televisão comercial do mundo cobrindo 99,55% do total da população brasileira. Tem como diferenciais o jornalismo (especialmente o Jornal Nacional e Fantástico) e a produção de telenovelas. Possui 120 afiliadas e inúmeros canais, incluindo plataformas de streaming e transmissão para outros países.

**Rede Record:** Fundada em 1993 pelo empresário Paulo Machado de Carvalho, a Rede Record é uma rede de televisão aberta que passa em canal aberto e tem como nicho conteúdo religioso. Está na segunda colocação de audiência nacional pelo IBOPE e possui um trabalho de digitalização bem avançado, já tendo uma plataforma de streaming e uma grande presença nas redes sociais.

**SBT:** O SBT foi fundado pelo empresário e comunicador Sílvio Santos em 1981 após o Grupo Silvio Santos ganhar uma licitação do Governo Federal para comprar a então “TV Tupi”. Hoje, ela ocupa a terceira posição na média da audiência nacional medida pelo IBOPE. Podemos considerar que o SBT ocupa dois nichos: o de programa de auditório, como programa do Ratinho e o próprio programa do Sílvio Santos, e o de programas infantis, como a novela *Chiquititas* e desenhos animados.

**Rede Bandeirantes:** A rede Bandeirantes foi criada em 1967 por João Jorge Saad após a compra da Rádio Bandeirantes. Hoje, a quarta posição na média nacional do IBOPE. Atua historicamente com esportes em geral e também com política, sendo tradição a produção dos primeiros debates de cada temporada de eleição no Brasil.

**TV Brasil:** É a emissora aberta oficial do Poder Executivo Brasileiro, além de ser obrigatoriamente transmitida em operadoras de TV paga. Atualmente, televisiona uma mistura de programas independentes, seriados e transmissão de eventos esportivos.

No que tange às tendências na indústria, a maior delas é a transformação digital. Com o advento de plataformas de streaming, a indústria televisiva tem se reinventado para disponibilizar sua grade na internet ao vivo e sob demanda. Um exemplo é a Globo Play, que oferece não só produções originais como também filmes e séries internacionais para competir com gigantes tais quais Netflix e Amazon Prime.

Além disso, tem-se o crescimento significativo da personalização no consumo de conteúdo. Através de algoritmos de predição, os principais streamings e redes sociais já conseguem recomendar o conteúdo mais relevante para cada usuário de modo certeiro. Assim, há grande pressão para que a indústria televisiva tradicional também agregue mais personalização em suas operações.

Ademais, o Brasil hoje oferece solo fértil para a produção de séries ficcionais e não ficcionais. Seguindo o sucesso desse modelo em outros países, existe significativa motivação para investir em documentários e minisséries nacionais, como demonstrado pelo sucesso de títulos tais quais “Irmandade” e “3%”.

## Análise aprofundada seguindo o modelo de 5 forças de Porter

**Ameaça de novos entrantes:** O estabelecimento de uma nova emissora de televisão exige grande capital inicial – com 70% de acionistas brasileiros –, uma concessão da Anatel e uma aprovação do Ministério de Comunicações quanto à sua proposta de programação e condições técnicas e financeiras. Sem dúvida, é uma empreitada que exige muitos recursos, paciência e competência. Além disso, uma vez estabelecida, uma nova emissora teria ainda o desafio de concorrer com as grandes redes de televisão brasileiras por audiência. Nesse sentido, percebe-se que há barreiras significativas para novos entrantes, de modo que a ameaça destes é baixa. No mercado atual existem grandes players, o que cria uma barreira de

novos entrantes com alta capacidade de crescimento. Desse modo, por mais que hoje a produção de conteúdo seja de fácil acesso, a distribuição (caso da rede gazeta) é um espaço com poucas empresas e que possuem quase um monopólio no segmento. Por fim, a possibilidade de novos entrantes é baixa, pois existe uma demanda de capital inicial, caixa e competência de desafiar grandes competidores nesse mercado.

**Poder de barganha dos fornecedores:** Os fornecedores são os produtores de conteúdo nacionais (incluindo roteiristas, diretores, atores, etc) e internacionais (distribuidores de filmes, séries e desenhos animados). Nesse sentido, tem-se um baixo poder de barganha no cenário brasileiro devido às poucas possibilidades de artistas televisionarem suas criações em emissoras de grande porte. Entretanto, quando se trata de empresas internacionais, o panorama muda, pois estas têm a possibilidade de vender para inúmeras emissoras, plataformas de streaming e consumidores diretos por todo o globo. Portanto, nesse caso, seu poder de barganha aumenta consideravelmente.

**Poder de barganha dos compradores:** O cliente, para a Rede Gazeta, pode ser tanto a “audiência” quanto os “patrocinadores”. Do lado da audiência, percebe-se um aumento de poder de barganha nos últimos anos devido à polarização político-ideológica que o Brasil tem enfrentado. Assim, telespectadores cobram atitudes condizentes com suas opiniões por parte das emissoras e tendem a migrar para outros canais quando não recebem o que esperam. Já do lado dos patrocinadores, a TV deixou de ser a primeira opção de praça para muitos. Com os avanços da tecnologia e dos algoritmos de marketing segmentado, existem hoje muitos outros meios de atingir clientes com maior eficácia e menor capital investido. Assim, os anunciantes passam a ter um maior poder de barganha para com as emissoras.

**Ameaça de serviços/produtos substitutos:** A indústria televisiva tem encontrado agressiva competição nas plataformas de streaming e redes sociais. Cita-se, por exemplo, Netflix e Youtube como importantes concorrentes, atraindo 33% e 64% da população brasileira atualmente. Entretanto, dado que, segundo a pesquisa do Kantar IBOPE, a televisão ainda persiste em 97% dos lares nacionais e teve seu consumo intensificado durante a pandemia, infere-se que a ameaça de substitutos é mediana, pois os dados mostram que ela ainda domina o momento de lazer do brasileiro médio.

**Rivalidade entre concorrentes:** Historicamente, emissoras de TV aberta foram sempre muito competitivas, lutando por certos artistas e punindo aqueles que migravam para uma rival. Entretanto, esse cenário tem se attenuado. Hoje, já se vê artistas e atores passando da Globo para o SBT, ou do SBT para a Record, sem grandes polêmicas. Uma exceção é o caso Globo-Record, as quais ainda não compartilham artistas e apresentadores. Ainda assim, é fato que a competição por direitos de transmissão, especialmente esportivos, é acirrada, assim como por scores de audiência absolutos.

#### 4.1.2. Análise SWOT

| MATRIZ SWOT   |   |   |
|---------------|---|---|
|               | Fatores Internos (Controláveis)   | Fatores Externos (Incontroláveis)   |
| Pontos Fortes | <b>Forças</b> <ul style="list-style-type: none"> <li>- Rede televisiva com maior alcance no Espírito Santo;</li> <li>- 16 veículos de comunicação que se conectam todos os dias com os capixabas;</li> <li>- Afiliada da Rede Globo;</li> <li>- Alta qualidade de equipamentos técnicos e de filmagem.</li> </ul>                                       | <b>Oportunidades</b> <ul style="list-style-type: none"> <li>- Popularização de serviços on-demand para conteúdo televisivo, abrindo espaço para investimentos em streaming de programas da TV Gazeta;</li> <li>- Baixa incidência de conteúdo culturalmente capixaba para a população local.</li> </ul>   |
| Pontos Fracos | <b>Fraquezas</b> <ul style="list-style-type: none"> <li>- Empresa tradicional;</li> <li>- Poucos funcionários focados na área de inovação;</li> <li>- Dependência de dados do People Meter, que não discriminam demográficos e grade de programação com acurácia satisfatória;</li> <li>- Limitados pelas decisões executivas da Rede Globo.</li> </ul> | <b>Ameaças</b> <ul style="list-style-type: none"> <li>- Alta concorrência no mercado de comunicação, intensificada pelos artistas de redes sociais;</li> <li>- Preferência por outros aparelhos eletrônicos e mídias sociais para entretenimento por partes dos telespectadores;</li> <li>- Crescimento das plataformas de streaming;</li> <li>- Enfraquecimento da mídia tradicional (em especial noticiários) por conta da polarização informacional do país;</li> <li>- Certo desinteresse da população brasileira por noticiários incisivamente políticos.</li> </ul> |

#### 4.1.3. Planejamento Geral da Solução

Para nosso projeto, recebemos um documento Excel com 18 abas, cada qual contendo taxas de audiência colhidas pelo PeopleMeter em diferentes períodos e divididas por gênero, faixa etária e classe socioeconômica.

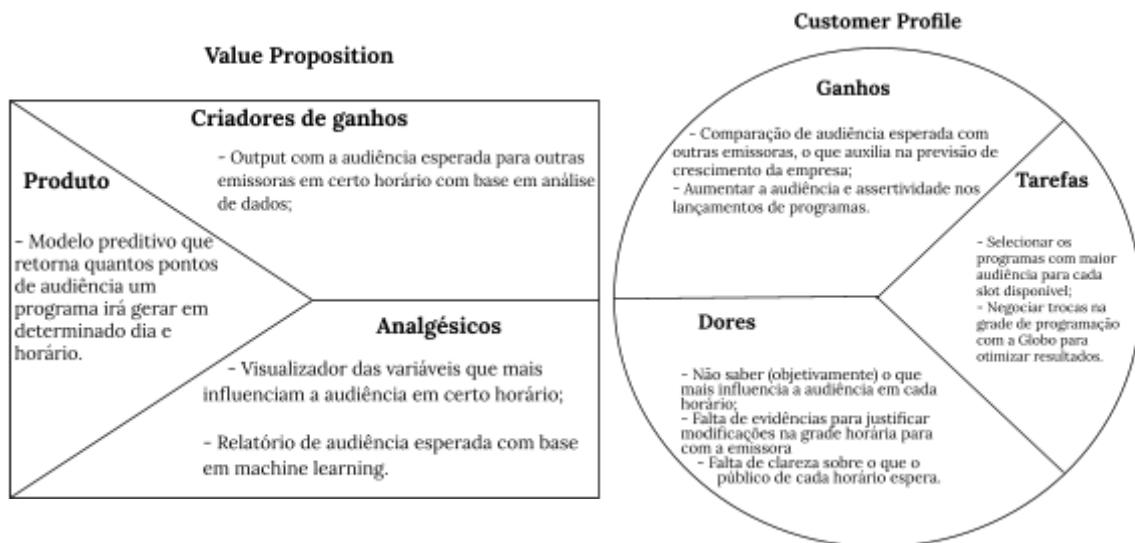
Com esse documento, pretendemos construir um modelo preditivo em que o usuário (funcionário da Rede Gazeta) possa inserir a data e horário desejados para um novo programa, assim como o gênero do programa em questão. Com isso, o objetivo é prever o score de audiência nessa data e slot, indicando as variáveis que mais tiverem peso para o modelo para, assim, criar ações mais efetivas para maximizar o retorno financeiro do evento.

Nesse sentido, nosso produto trata de regressão, pois nós vamos retornar um valor dentre infinitas opções. O output secundário, no entanto, relacionado às variáveis utilizadas na predição, talvez seja classificatório. Ainda não temos como saber.

Além disso, em termos práticos, a solução proposta será utilizada pelas equipes de Programação e Marketing da Rede Gazeta. Desse modo, de acordo com os dados fornecidos por esse modelo, a Rede Gazeta pode definir um esforço (maior ou menor) necessário para divulgar certo programa, realizar trocas na grade de horário ou mesmo negociar slots com a Globo. Tudo isso, se bem aplicado, contribuirá para um maior retorno financeiro para os investimentos da empresa.

Por fim, o critério de sucesso ainda não foi bem definido, dado que não começamos a implementar o modelo. Entretanto, antecipamos que ele será determinado de acordo com uma tolerância máxima para o erro da predição. Isso, por sua vez, será aferido na fase de testes com um training set derivado dos dados que recebemos. Logo, consideramos o modelo preditivo um sucesso ele retorna uma margem de erro pequena/desconsiderável.

#### 4.1.4. Value Proposition Canvas



## 4.1.5. Matriz de Riscos

| Matriz de Risco |   |             |       |  |   |   |  |  |       |                                     |             |
|-----------------|---|-------------|-------|--|---|---|--|--|-------|-------------------------------------|-------------|
| Probabilidade   |   | Riscos      |       |  |   |   | Oportunidade   |  |       |                                     |             |
| Muito Alta      | 5 |             |       | Atrasos por conta de membros do grupo faltando                             |   |   |  |  |       |                                     |             |
| Alta            | 4 |             |       |  |   |   | Entregar análises fidedignas considerando as capacidades técnicas do grupo | Realizar os autoestudos antes das aulas e desenvolvimento s relacionados |       |                                     |             |
| Médio           | 3 |             |       | Parceiros de projeto com expectativa diferente de produto, superestimando. | Concentração de tarefas em pessoas específicas do grupo                   | Nosso produto não se destacar em relação aos outros grupos    | A solução ser adotada pela Rede Gazeta                                     |  |       | Expandir a audiência da rede Gazeta |             |
| Baixa           | 2 |             |       | Não conseguir entregar o projeto a tempo                                   | Não conseguir analisar todos os dados e trazer vieses no modelo preditivo |   |  |  |       |                                     |             |
| Muito Baixa     | 1 |             |       | Grupo não se entrosar  | Mudanças na LGPD, impactando as amostras que o aparelho pode coletar      | Violação dos dados do ibope Não entregar previsões assertivas |  |  |       |                                     |             |
|                 |   | 1           | 2     | 3  | 4   | 5   | 5  | 4  | 3     | 2                                   | 1           |
|                 |   | Muito Baixo | Baixo | Médio  | Alta  | Muito Alta  | Muito Alta   | Alta   | Médio | Baixo                               | Muito Baixo |
| Impacto         |   |             |       |  |   |   |  |  |       |                                     |             |

 Matriz de risco - Grupo 5

#### 4.1.6. Personas

##### Persona 1



Marta Lopes, 32 anos, Analista de Comunicação.

*"Comunicativa, gosta de trabalhar com as pessoas, viajar e conhecer diferentes culturas"*

*Biografia:* Migrou do Marketing para a área de Comunicação da Rede Gazeta; Decidida, prática e ótima negociadora.

*Dores/Motivações atuais com o problema:* Não consegue negociar com os anunciantes de forma que correlacione o horário com a classe social que é alvo do anúncio.

*Objetivos/necessidades específicas em relação ao problema:* Focalizar os anunciantes para horários com maior potencial de compradores; Ter evidências para negociar taxas maiores com anunciantes fora do tradicional "horário nobre".

##### Persona 2



Giovanna Mattos, 28 anos, Gerente Geral de Marketing.

*"Analítica, adora dados e quer transformar o ambiente em que trabalha através da tecnologia."*

*Biografia:* Recentemente contratada e ainda se familiarizando com a empresa e suas particularidades; Experiência com Google Ads e análise de dados; Trabalhou como freelancer por bastante tempo.

*Dores/Motivações atuais com o problema:* Não há a possibilidade de prever a necessidade de reforço de campanhas de marketing direcionadas à horários; Falta de evidências para aprovar o orçamento de marketing.

*Objetivos/necessidades específicas em relação ao problema:* Reavaliar a divulgação de novos eventos que tiveram um menor potencial de audiência; Identificar fatores que mais contribuem para a audiência de um programa a fim de enfatizá-los na campanha de marketing; Ter evidências para negociar orçamentos maiores e promoções para o departamento de marketing, a partir da comparação da audiência esperada e audiência consolidada após campanhas.

### Persona 3



Rodrigo Souza, 35 anos, Gerente de Operação e Programação.

*"Funcionário de longa-data desde o período de estágio. Enfrenta dificuldades para justificar os seus planos de ação de forma objetiva desde cedo."*

*Biografia:* Funcionário de longa-data muito familiarizado com a grade atual; Opiniões fortes sobre os programas existentes e muitas ideias de melhoria; Intuitivo, tende a tomar decisões com base em suas impressões subjetivas.

*Dores/Motivações atuais com o problema:* Não saber (objetivamente) o que mais influencia a audiência em cada horário; Falta de evidências para justificar modificações na grade horária para com a Globo; Falta de clareza sobre o que o público de cada horário espera.

*Objetivos/necessidades específicas em relação ao problema:* Selecionar os programas com maior audiência para cada slot disponível; Identificar o que faz um programa ter sucesso em cada slot para produzir melhores eventos; Negociar troca de programação em certos slots com a Globo; Suprir expectativas do público de determinado slot.

## 4.1.7. Jornadas do Usuário



**Rodrigo, Gerente de Operação e Programação**

**Cenário:** Rodrigo quer realizar um plano de ação mais objetivo para a audiência de um novo programa.

| <b>Expectativas</b>  |   |   |  |   |
|--|---|---|--|---|
| <b>FASE 1<br/>(Organização de dados)</b>   | <b>FASE 2<br/>(Análise)</b>   | <b>FASE 3<br/>(Deduzindo)</b>   | <b>FASE 4<br/>(Preparando)</b>   | <b>FASE 5<br/>(Conferir e Estabelecer)</b>  |
| <p>1.Começa a organizar (quase que manualmente) os números de audiência, características dos telespectadores e horários que possam se relacionar com o novo programa;</p> <p>2. Para acessar esses dados (tabelas em Excel) precisa pedir acesso para outros setores.</p> <p><b>'É essencial organizar todos os dados que possam se relacionar com o contexto de um novo programa, mesmo o processo sendo extremamente cansativo'</b></p>  | <p>1.Confere mais cuidadosamente os dados selecionados, com viés de análise;</p> <p>2.Checa o número de audiência, características do público e horário de antigos programas que se relacionam com o novo;</p> <p>3.Imagina o quanto bem esse novo programa poderia ser encaixado com base em experiências anteriores.</p> <p><b>'Pelo o que já se passou até hoje, como seria o desempenho de um novo programa nesse cenário? Imagino o quanto bem esse programa e nossa rede possa se tornar '</b></p>  | <p>1.A partir do que se foi imaginado e concluído (tendendo ao subjetivo), temos as primeiras propostas de encaixe da nova programação expostas, a partir da audiência;</p> <p>2.Discorre as novas propostas com o que foi formulado.</p> <p><b>'Bom, contudo, seria ótimo ter argumentos mais sólidos para a implementação de um novo programa na grade. Ainda tenho minhas dúvidas em relação a essas novas propostas'</b></p>  | <p>1.Organiza os dados coletados anteriormente para serem inseridos no modelo preditivo;</p> <p>2.Apresenta as circunstâncias e dúvidas trabalhadas para o modelo preditivo.</p> <p><b>'Inserir todos os dados no modelo preditivo é bem cansativo/chato'</b></p>  | <p>1.Compara as suposições subjetivas já feitas com aquelas desenvolvidas no modelo preditivo, fazendo questionar o que se foi concebido anteriormente ou apenas reforçando;</p> <p>2.Com esse estudo mais completo, sugestões dessa nova grade de horário são apresentadas.</p> <p><b>'Agora sim, embasado e completo. Estou satisfeito com o resultado do modelo preditivo!'</b></p>  |

### Oportunidades

- Automatizar e deixar mais rápida a organização inicial de dados;
- Deixar mais prática a inserção de dados no modelo preditivo.

miro



**Giovanna Mattos, Gerente Geral de Marketing**

**Cenário:** Giovanna quer identificar exatamente (ou o mais próximo disso) o melhor horário para divulgar comerciais e propagandas dos novos programas.

| <b>Expectativas</b>  |  |  |   |   |
|--|--|--|---|---|
| <b>FASE 1<br/>(Destrichando/Organizando)</b>   | <b>FASE 2<br/>(Análise)</b>  | <b>FASE 3<br/>(Deduzindo)</b>  | <b>FASE 4<br/>(Preparando)</b>  | <b>FASE 5<br/>(Conferir e Estabelecer)</b>  |
| <p>1.Após receber as informações do novo programa, confere quais programações já estabelecidas mais se assemelham a ele;</p> <p>2.Pede acesso aos dados (tabelas Excel) desses programas semelhantes (horário, público atingido, antiga forma de divulgação) para outros setores.</p> <p><b>'Esse é um trabalho muito manual e burocrático, tenho que pedir diversas tabelas e informações para outros setores do trabalho '</b></p>  | <p>1.Confere mais cuidadosamente os dados selecionados, com viés de análise;</p> <p>2.Checa em quais programas e horários o público alvo está mais ativo;</p> <p>3.Compara como antigas divulgações desse gênero de programa foram feitas e seus impactos, mantendo acertos e evitando erros.</p> <p><b>'Gosto muito de poder exercer esse meu lado analista, mas um olhar pessoal ainda é necessário'</b></p>  | <p>1.A partir do que se foi imaginado e concluído na análise, temos as primeiras propostas de encaixe da divulgação do novo programa;</p> <p>2.Discorre as novas propostas com o que foi formulado.</p> <p><b>'Bom, mas ainda é algo um tanto que subjetivo, queria ter ainda mais exatidão nessa minha proposta de marketing em divulgação'</b></p>  | <p>1.Organiza os dados coletados anteriormente para serem inseridos no modelo preditivo;</p> <p>2.Apresenta as circunstâncias e dúvidas trabalhadas para o modelo preditivo.</p> <p><b>'Inserir todos os dados no modelo preditivo é bem cansativo/chato'</b></p>  | <p>1.Compara as suposições subjetivas já feitas com aquelas desenvolvidas no modelo preditivo, fazendo questionar o que se foi concebido anteriormente ou apenas reforçando;</p> <p>2.Com esse estudo mais exato e completo, a forma de divulgação desse novo programa é apresentada.</p> <p><b>'Agora sim, ainda mais próxima da exatidão. Estou satisfeita, com o resultado do modelo preditivo!'</b></p>  |

### Oportunidades

- Deixar mais prático e evitar intermediários na organização inicial de dados;
- Tornar a inserção de dados no modelo preditivo menos manual e mais rápida.

Obs: Caso esteja com dificuldade de visualização, a imagem está com dimensões maiores na seção Anexos.

## 4.2. Compreensão dos Dados

Recebemos duas planilhas em formato XLSX que agrupam dados coletados pelo Kantar IBOPE Media a partir de um aparelho de depuração de audiência chamado “People Meter”. Esse aparelho torna possível registrar o perfil de quem está assistindo com os demográficos de que faz parte, o número total de indivíduos atingidos e o tempo de consumo.

O primeiro arquivo, denominado “TV\_Histórico”, contém 18 abas elencando dados de audiência por emissora e dia da semana. Mais especificamente, cada emissora é dividida em três abas: uma para dias úteis (156.673 linhas), uma para sábado (31.105 linhas) e uma para domingo (também 31.105 linhas). O período analisado, por sua vez, é de 6 de junho de 2020 a 25 de junho de 2022, sendo que cada dia é tabelado de cinco em cinco minutos. Uma peculiaridade, nesse sentido, é que o horário a cada dia começa em 24:00:00 (correspondendo à meia-noite) e vai até 29:55:00 (correspondendo às 5:55:00 da manhã); depois, a contagem se reinicia às 6:00:00.

Cada aba contém as porcentagem de três métricas de audiência: Rating (considerado o score de audiência), Fidelidade (parcela dos televisores que ficou no mesmo canal por pelo menos um minuto), Reach (parcela de televisores alcançada) e Share (parcelas de televisores sintonizados com certa emissora dentre todos os ligados em dado momento). Esse valores são, por sua vez, subdivididos em classe socioeconômica (AB, C1, C2 e DE), sexo (masculino e feminino) e idade (4-11 anos, 12-17 anos, 18-24 anos, 25-34 anos, 35-49 anos, 50-59 anos e 60+ anos).

Já o segundo arquivo, “grade\_Diária\_06\_2020\_a\_06\_2022.xlsx”, traz uma coluna “Praça”, correspondente ao local de transmissão, “Data”, “Faixa Horária” e uma coluna para cada emissora, contendo a programação para cada slot de junho de 2020 a junho de 2022.

Para fins de análise de dados, esse documento foi agregado com o primeiro através da adição de duas colunas (“Programa” e “Gênero”) seguindo a paridade de horário entre as tabelas. Além disso, os horários foram formatados em intervalos de 30 minutos através de média aritmética para repetição. O campo de data também foi dividido em colunas de ano, mês e dia.

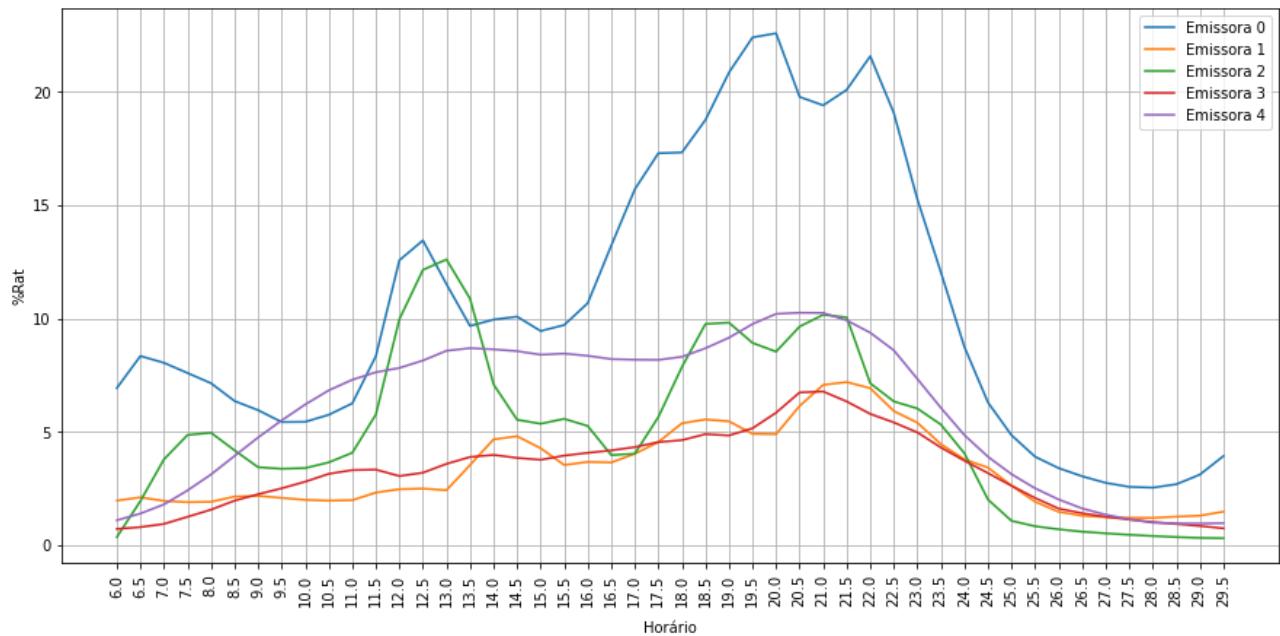
Ademais, é importante notar que não há contingências em relação à qualidade e preservação dos dados, ou seja, células vazias ou “NaN” não são um problema até o momento.

Por fim, os dados disponibilizados pelo Kantar IBOPE são de natureza sigilosa, de modo que o nome das emissoras não deve ser mencionado em qualquer documento público.

## Análise preliminar dos dados

Para todas as análises, foram consideradas apenas as audiências em Rat, conforme recomendação do cliente. Isso se dá porque o Rat é justamente um cálculo mais significativo sobre os valores de Fid e Shr.

### Audiência por hora por emissora (Seg a Sex)



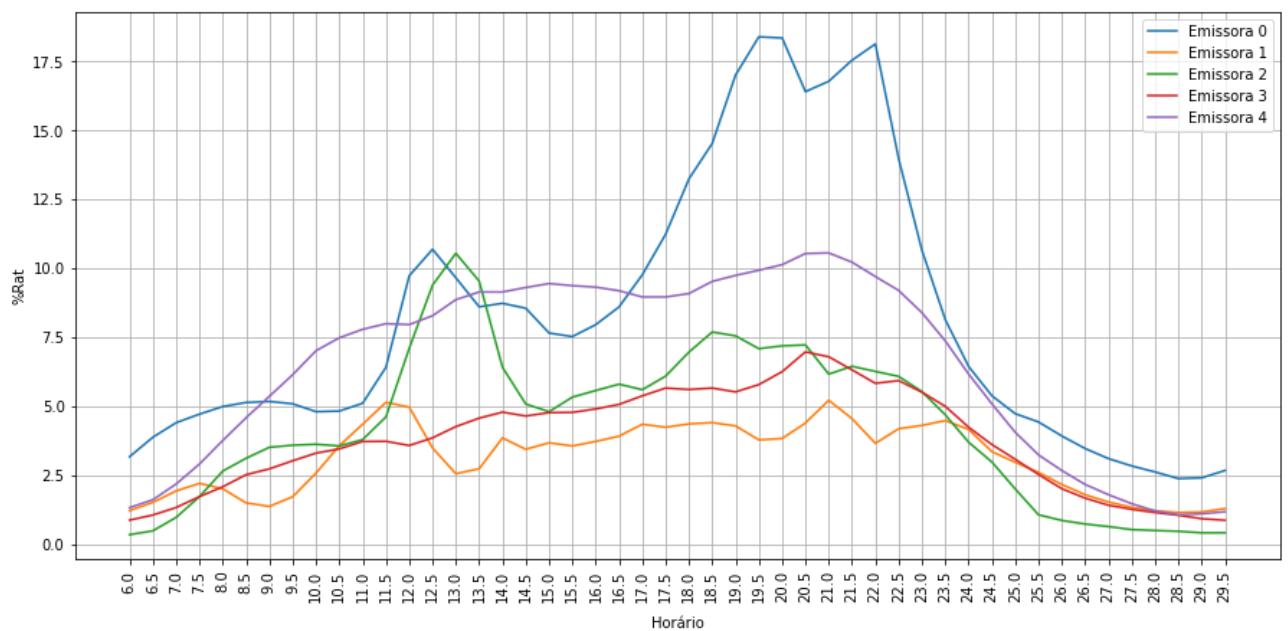
Este gráfico apresenta cada emissora como uma linha e mostra a audiência de Rat de domicílios totais a cada hora do dia durante a semana. Nesse sentido, há uma dominância clara da Emissora 0 durante todo o período, exceto das 9h30 às 11h15, quando a Emissora 4, representando itens não identificados, como streaming, a supera. Uma hipótese que explica esse comportamento é o fato de este ser o horário preferido por donos(as) de casa para realizar tarefas domésticas; nesse contexto, é provável que elas utilizem os desenhos animados disponíveis em plataformas de streaming para distrair as crianças, segundo insight do parceiro de projetos.

Ademais, o primeiro pico relativo ocorre às 6h30. Sugere-se que isso ocorra devido à veiculação do primeiro evento jornalístico do dia justamente no momento em que a maioria das pessoas acorda para ir à escola ou ao trabalho. Isso volta a ocorrer ao meio-dia, aproveitando a atenção das pessoas em seu horário de almoço. Aqui, há uma competição acirrada entre a Emissora 0 e a Emissora 2, dado que ambas apresentam o segundo jornal do dia nesse horário. Percebe-se, portanto, que a audiência se divide entre as duas, o que indica a necessidade de se diferenciar e segmentar o conteúdo a fim de melhor atender a população capixaba. Uma possibilidade seria dedicar mais tempo a assuntos locais e temas culturais para gerar mais identificação no público, instigando um sentimento de orgulho e pertencimento que os fará voltar para o programa dia a dia.

A partir daí, há uma baixa de audiência durante o final do bloco jornalístico e bloco de filmes, provavelmente por ser um horário em que a maioria das pessoas está ocupada com suas tarefas diárias, profissionais ou não.

A audiência volta a subir no final da tarde com a transmissão das primeiras telenovelas e aumenta consistentemente até cerca das 20h, quando começa o principal programa jornalístico do dia. Nesse momento, há uma queda súbita com grandes chances de ser motivada por divergências ideológicas. Por fim, o último pico se dá às 22h30, em que ora passam filmes, ora reportagens, ora reality shows populares.

### Audiência por hora por emissora (Sábado)



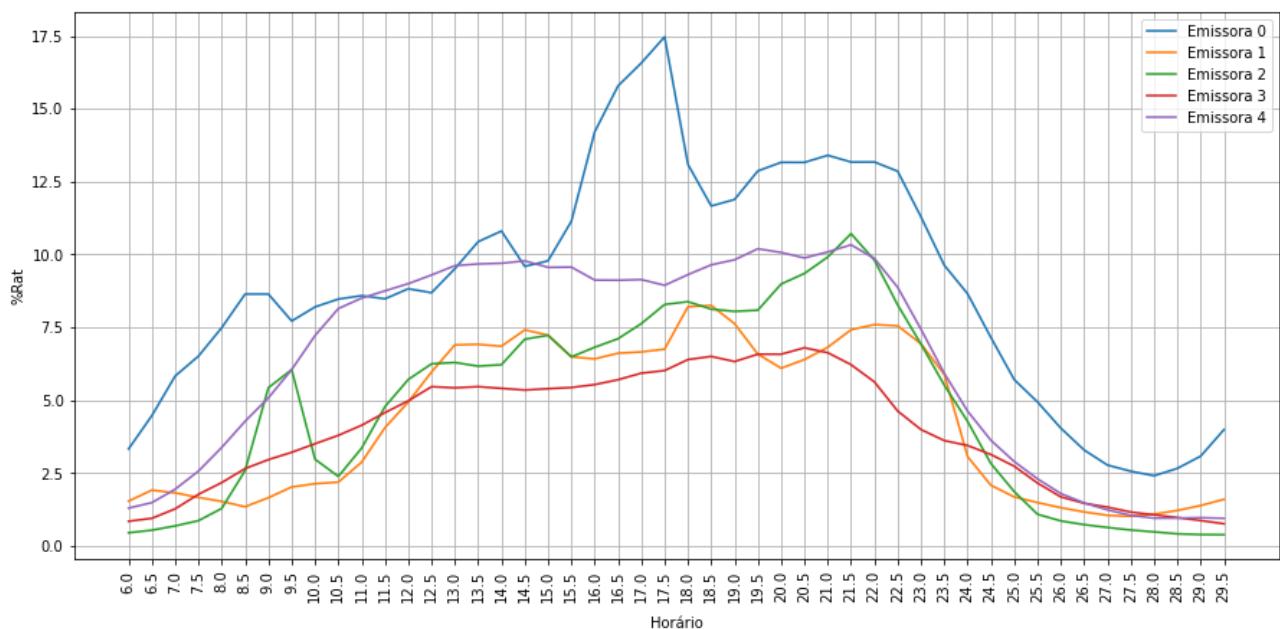
Aos sábados, nota-se uma configuração relativa semelhante de audiência entre as emissoras. Porém, ao tratar de valores absolutos, percebe-se que as porcentagem de Rat são claramente menores aos sábados, chegando à máxima geral de 18 contra a de 21 durante a semana.

Ademais, dentre outras particularidades, há uma presença considerável da Emissora 4, representando plataformas de streaming e outros conteúdos não identificados pelo PeopleMeter. Isso é visível pela primeira vez das 8h45 até as 11h30. Nossa hipótese, nesse caso, segue a linha da anterior em que pais e guardiões utilizam essas ferramentas para distrair as crianças durante o período da manhã. Outra explicação é um maior consumo de entretenimento durante o período da manhã em finais de semana, em que a família se reúne para tomar café e pode compartilhar conteúdos no tablet e no celular. Já os picos na hora do almoço seguem a mesma lógica do resto da semana, ainda que com valor absoluto inferior a 20%.

Durante a tarde, a Emissora 4 novamente sobe em audiência comparativamente, apesar de se manter no mesmo valor absoluto dos dias úteis. Isso indica que as pessoas assistem a conteúdo de streaming nesse horário mais consistentemente do que à TV aberta, isto é, em seu tempo de descanso, elas preferem streaming aos filmes e shows de auditório característicos dos sábados.

Por fim, o resto do dia segue o comportamento de segunda a sexta.

### Audiência por hora por emissora (Domingo)



Domingo é um dia atípico, como pode ser claramente visto pelas flutuações nas linhas. O dia já começa com audiência considerável e consistente das 8h30 às 9h para a Emissora 0, que veicula eventos rurais nesse período. Um fato curioso é que a queda da Emissora 0 entre 9h e 9h30 se alinha quase que perfeitamente com a subida da Emissora 2, indicando uma migração dos telespectadores para assistir à programação mais decididamente capixaba dessa emissora para esse slot.

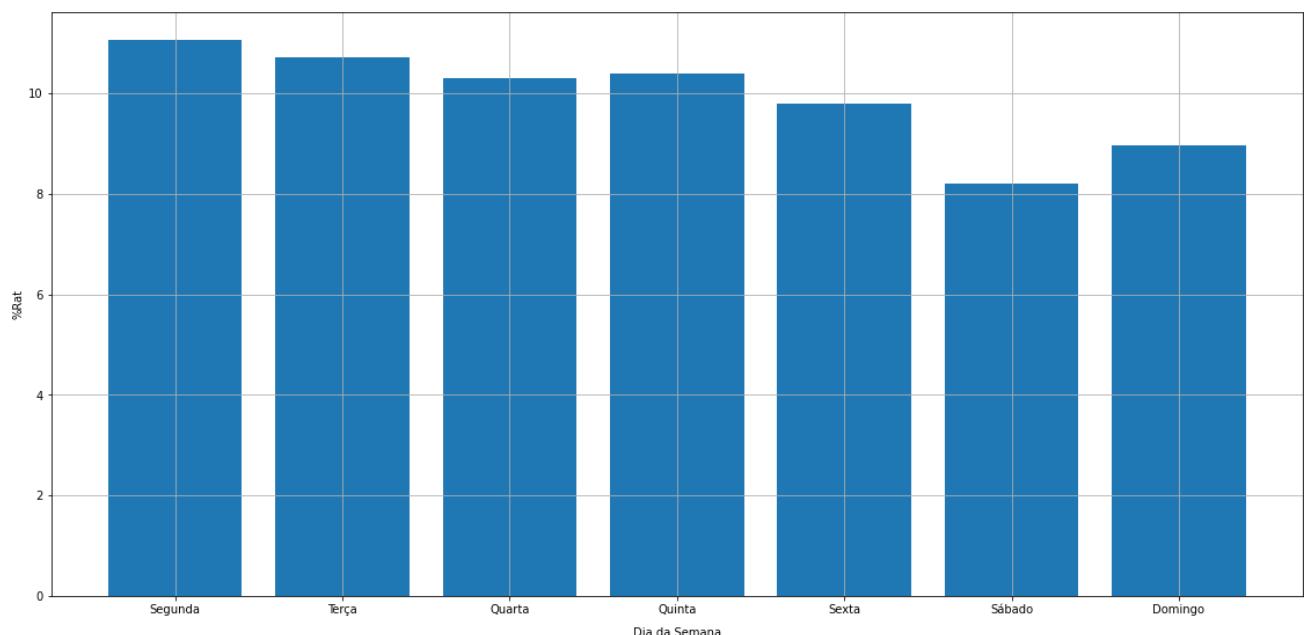
No horário do almoço, os picos não ocorrem mais devido à ausência de eventos jornalísticos. Ainda assim, o leve aumento às 13h30 para a Emissora 0 mostra que a população gosta de sentar para assistir a filme – conteúdo comum para esse período aos domingos – após se alimentar.

O pico principal acontece mais cedo, às 17h30, quando começam os programas de auditório mais populares de cada emissora. Além disso, existe um público cativo consistente para a programação das 20h às 23h, quando são transmitidos blocos de reportagens e reality

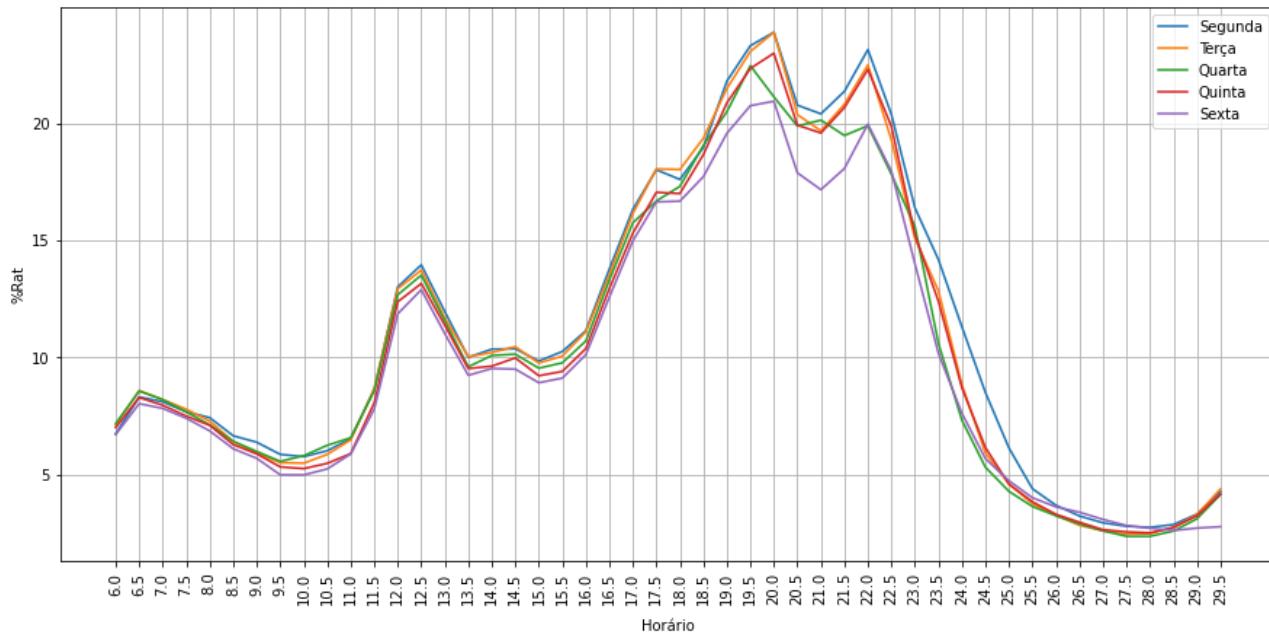
shows. Ainda assim, em segundo nível, a Emissora 4 ainda alcança uma audiência competitiva com seus eventos de auditório mais tradicionais.

### Audiência da Emissora 0 por dia da semana

Dada a dominância geral da Emissora vista nos gráficos anteriores, decidimos seguir as análises apenas com a audiência dela, pois outras emissoras não apresentam valores significativos em comparação.



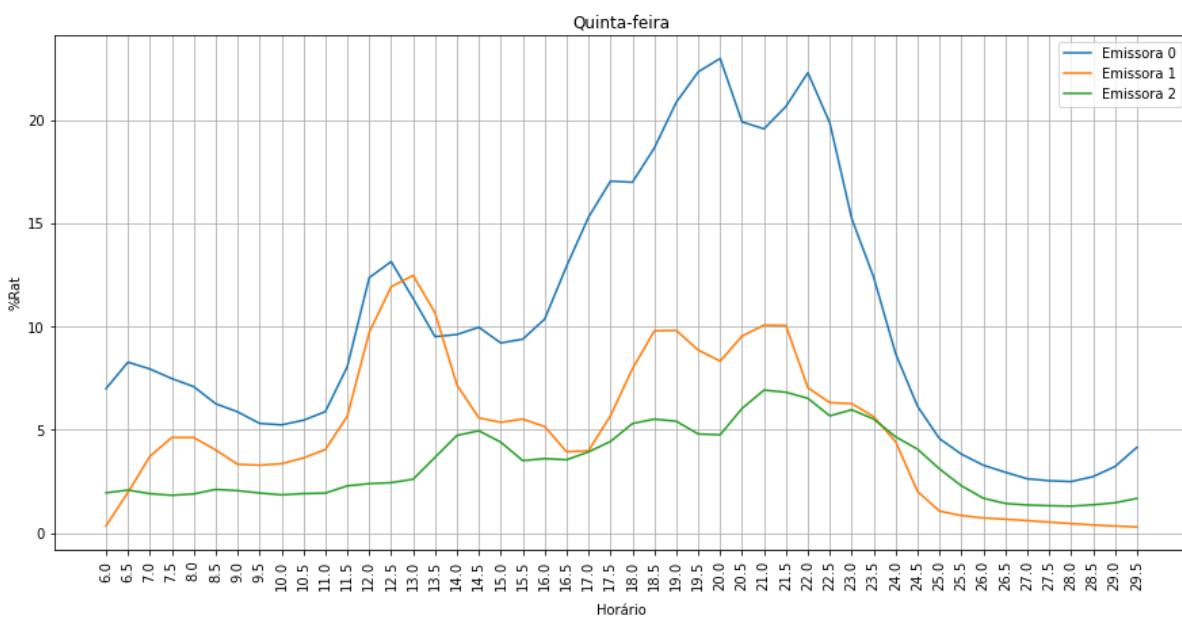
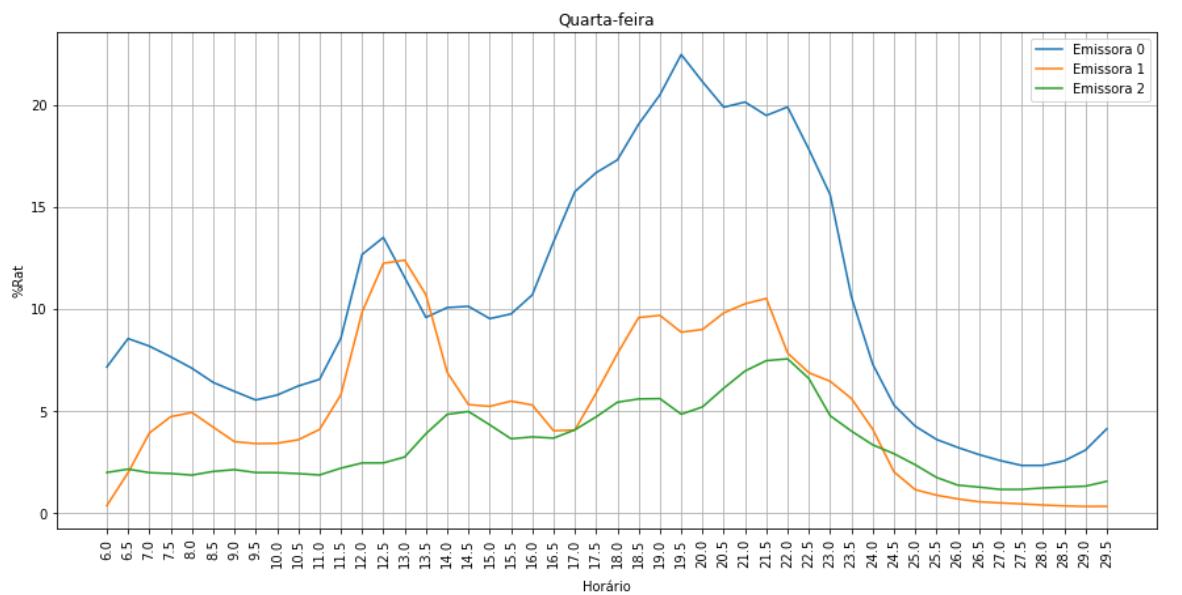
A primeira coisa a se notar é que as audiências para cada dia, exceto sábado e domingo, apresentam pouca variabilidade, ficando sempre em torno de 10-10.5 Rat%. Nesse sentido, as baixas no final de semana são fáceis de explicar: as pessoas têm mais tempo livre nesses dias e buscam outras formas de lazer além da televisão. Por outro lado, explicar por que mais pessoas sintoniza às segundas tem se provado um desafio. A programação semanal em certo mês tende a se manter regular, de modo que não grandes mudanças entre segunda e terça, por exemplo. Para aumentar nosso entendimento nesse aspecto, geramos também os gráficos de horário para cada dia da semana, a fim de visualizar quais blocos e, consequentemente, quais programas mais atraíam os telespectadores.



Como sugerido, as audiências são muito semelhantes durante a semana. Ainda assim, olhando com mais atenção, percebe-se que os valores para sexta-feira tendem a cair cada vez mais conforme o dia progride, decerto porque as pessoas preferem sair após o trabalho nessas noites em antecipação do final de semana.

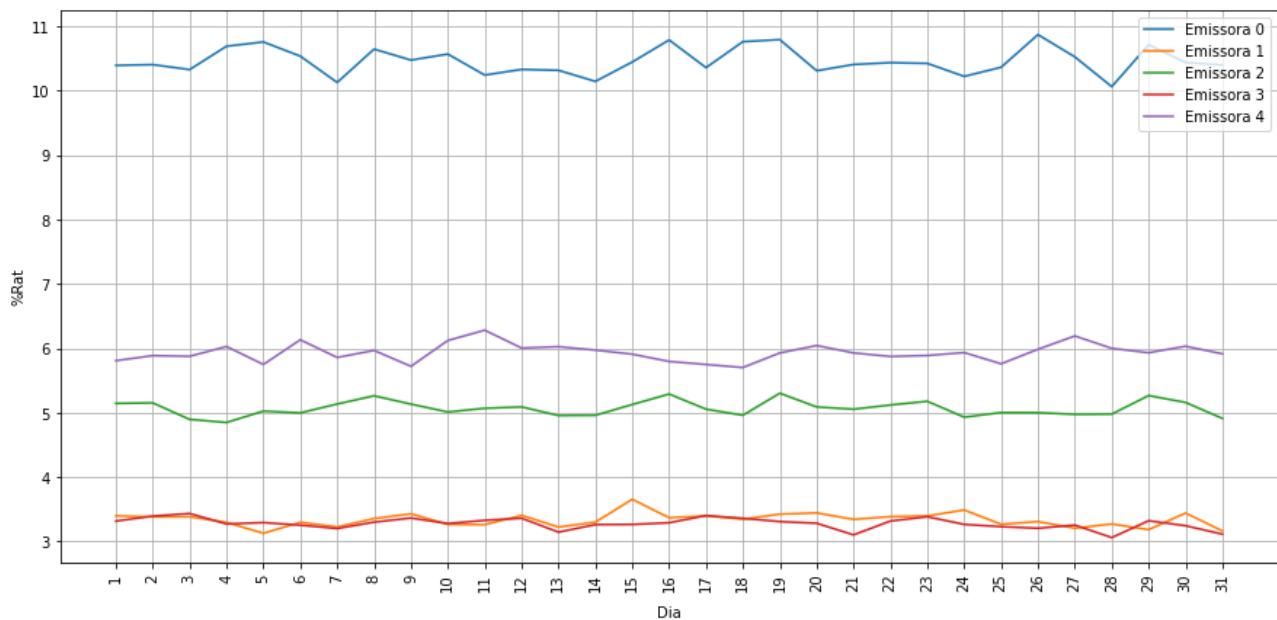
Já a terça-feira segue o padrão de segunda com muita precisão, exceto no final da noite e início da madrugada. Um ponto curioso aqui é que, pelo jeito como as horas são apresentadas no gráfico, temos a impressão de que a hora 29.0 no gráfico de segunda-feira representa a madrugada de terça-feira. Porém, na realidade, esse horário se refere à madrugada entre domingo e segunda-feira, pois, na planilha, todos os dias iniciam com horário 24.0 e só depois retornam para 6. Com esse insight, percebemos que um motivo para a maior audiência na madrugada de segunda-feira é simplesmente o fato de esta ser uma “extensão” de domingo, de modo que a muitas pessoas ainda não voltaram a seu horário de sono habitual e ficam acordadas até mais tarde.

Ademais, para examinar as baixas relativas nas quartas e quintas, plotamos também as audiências por horário para cada um desses dias considerando também outras emissoras. O objetivo era identificar se havia migração por algum programa específico.



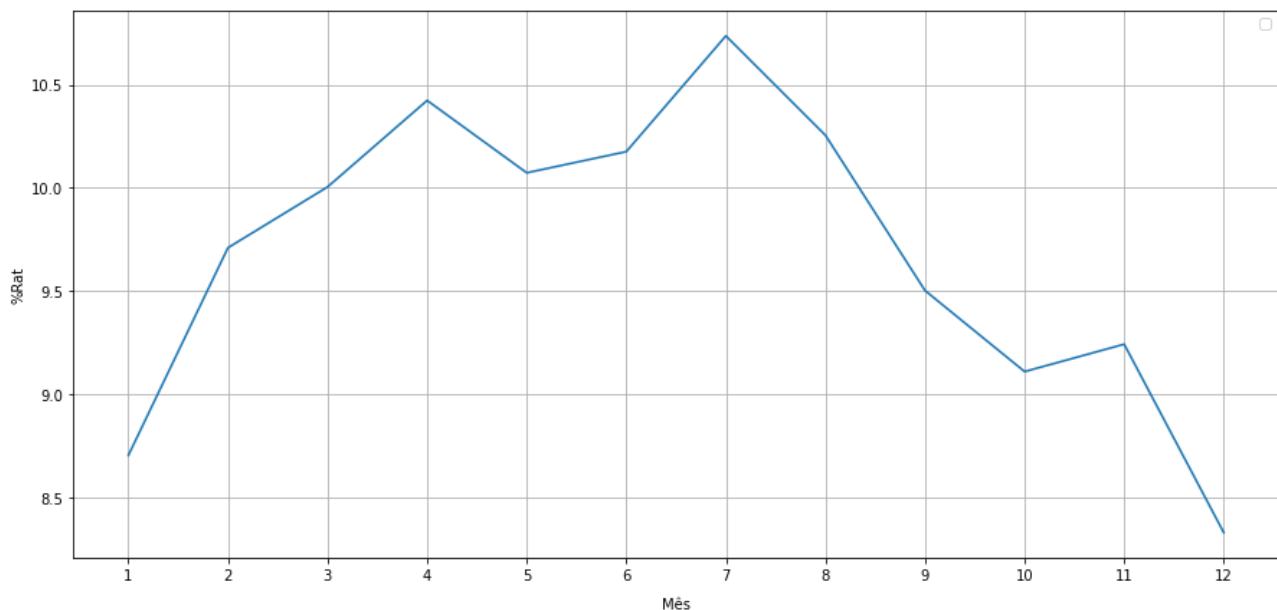
Ambos os dias, entretanto, apresentam flutuações semelhantes às dos outros dias da semana, como aferido pela média semana no primeiro gráfico mostrado nessa análise. Assim, ainda nos falta informação para induzir especificamente o que pode influenciar esse comportamento diferenciado às quartas-feiras e, em menor grau, às quintas.

## Audiência por dia do mês por emissora



Novamente, a dominância da Emissora 0 é clara, e as audiências flutuam em um intervalo muito pequeno. Nesse sentido, ao comparar as datas dos picos relativos com a programação vigente, há resultados inconclusivos. Em alguns casos, realmente houve eventos especiais, como jogos de futebol, reality shows e blocos de reportagens populares. Em outros, no entanto, não há explicação clara do porquê esse dia em particular atrairia mais público. Por isso, deduzimos que, salvo raras exceções, não há muita variação de comportamento com o passar do mês.

## Audiência da Emissora 0 por mês

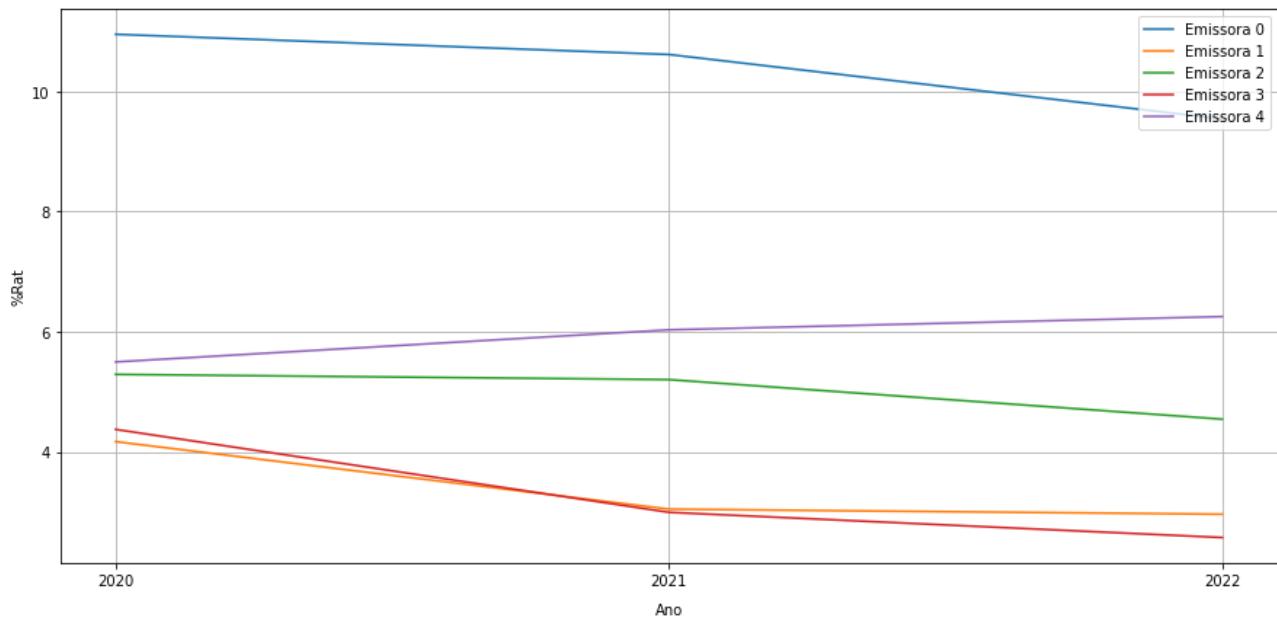


Há dois picos importantes na análise mensal de audiência. O primeiro ocorre em abril, provavelmente porque esse foi o mês em que o reality show mais popular da Emissora 0 foi televisionado nos últimos dois anos. Já o segundo se dá em julho, decerto pela veiculação das últimas Olimpíadas nesse período.

A queda que se segue pode ser explicada, em parte, pela usual troca de novelas e séries no segundo semestre. Nesse contexto, vê-se que a audiência ainda se mantém relativamente alta no mês de agosto, talvez pelo interesse da população nas novidades, mas diminui com o passar dos meses quando essa novidade inicial se esvai.

Há um leve aumento em novembro, que nós atribuímos como hipótese à finalização de atividades profissionais e escolares. É o momento em que as responsabilidades diminuem e as pessoas começam a tirar férias. Já em dezembro, com os planejamentos para as festas e possíveis viagens e outras formas de lazer presencial, a audiência volta a cair.

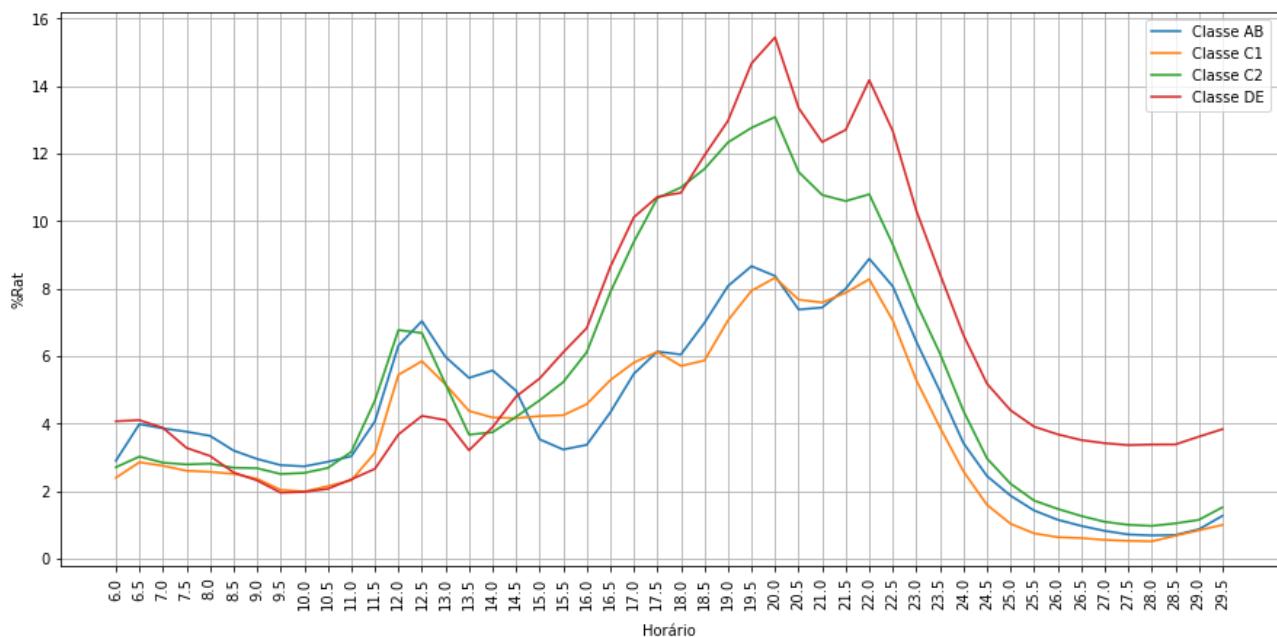
## Audiência por emissora por ano



Nota-se uma leve queda de audiência da Emissora 0 de 2020 para 2022. Isso é explicado pela normalização pós-pandemia. Durante a crise de COVID-19, as pessoas ficaram muito mais tempo em casa e, portanto, tiveram mais oportunidades de assistir à televisão, além de terem mais interesse nas notícias quanto ao status da doença. Com a reabertura comercial, os níveis de audiência voltaram para o que acreditamos serem níveis pré-pandemia. Entretanto, como não temos dados de anos anteriores, não podemos validar essa hipótese.

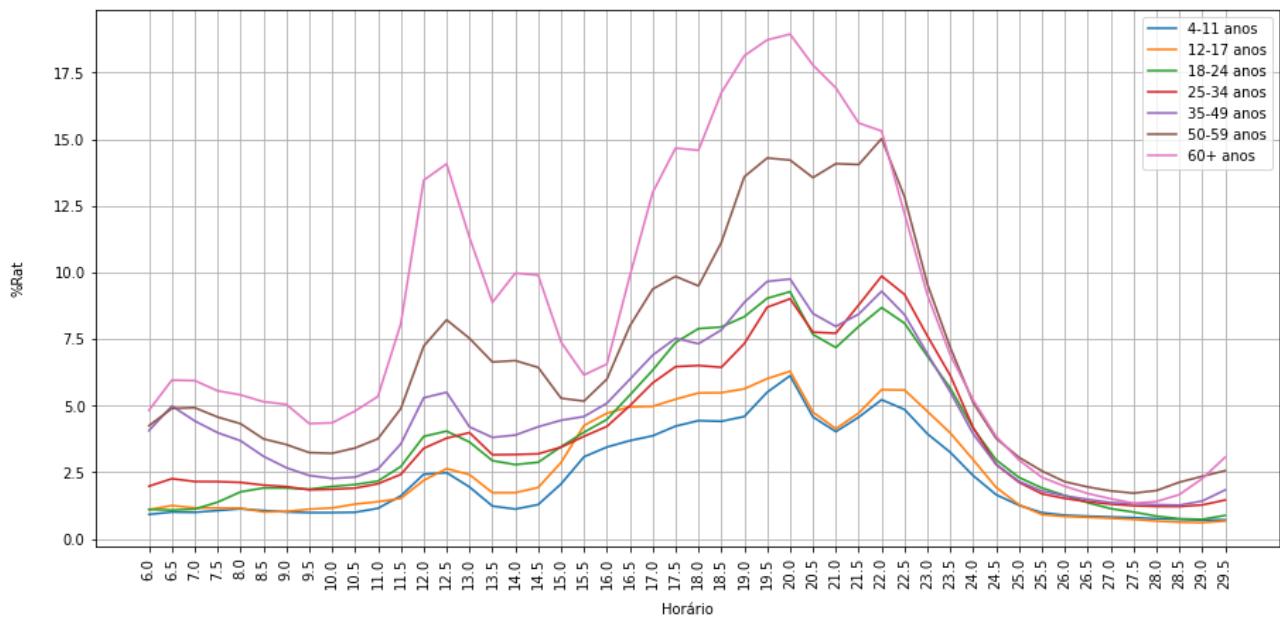
Percebe-se também que a Emissora 4, representando plataformas de streaming, teve certo crescimento nos últimos anos. Isso é corroborado por pesquisas que mostram que o consumo de streaming aumentou em 34% com a criação de mais conteúdo exclusivo ("State of Mobile 2022 - data.ai", 2022).

## Audiência da Emissora 0 por classe socioeconômica



Na análise socioeconômica, o pico geral pertence à classe DE no bloco jornalístico do período da noite. Assim, deduz-se que essa classe tem preferência pelo jornalismo tradicional, enquanto as outras classes consomem uma programação mais variada e menos política, com programas de auditório, jornais locais e novelas, durante a manhã e tarde.

## Audiência da Emissora 0 por faixa etária



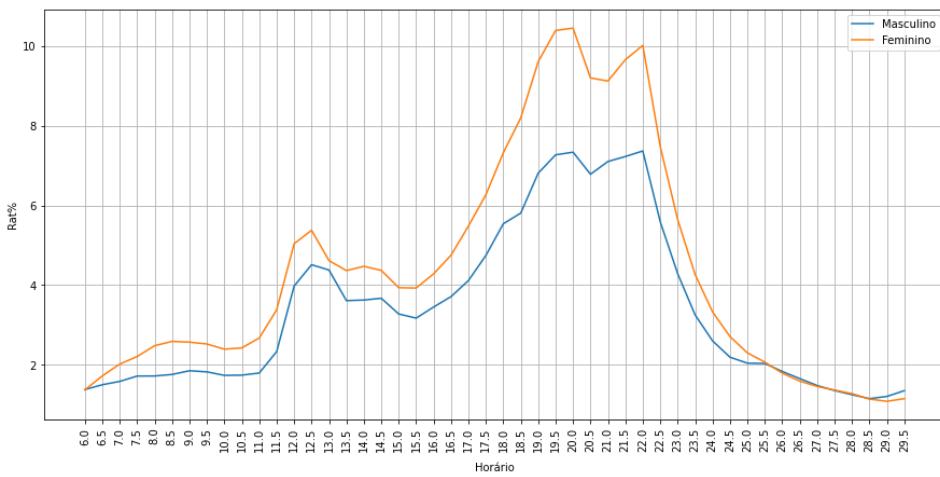
No que tange à faixa etária, a terceira idade domina em quase todos os casos e especialmente durante a manhã, provavelmente pela incidência de aposentadoria e horários mais flexíveis, além da tendência a acordar mais cedo (AMANDA CHAN 12 APRIL 2011, [s.d.]). Apesar disso, todas faixas apresentam comportamento e flutuações parecidas, ainda que em níveis diferentes, indicando que a categoria dos eventos a cada faixa horária agrada a população em geral de forma consistente. Um ponto interessante é que crianças aparecem com muito menos frequência, provavelmente pela escola e preferência por Youtube e outras plataformas.

## Audiência total da Emissora 0 por gênero

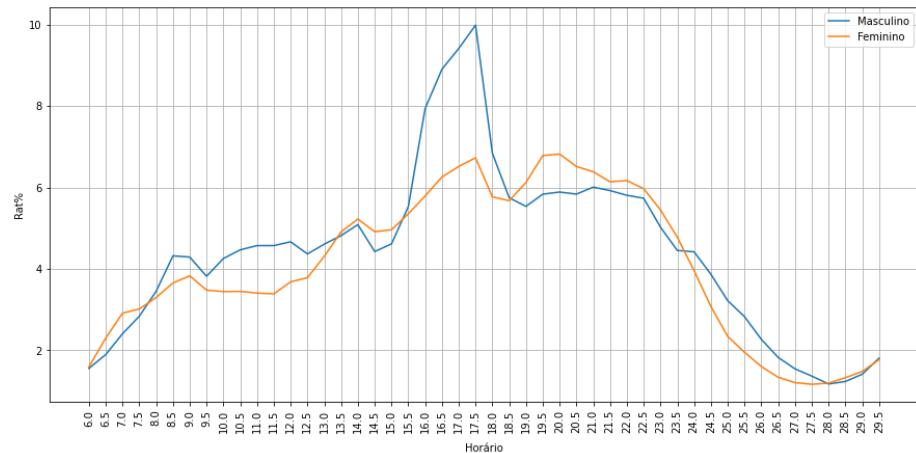
**Segunda a sexta**



**Sábado**

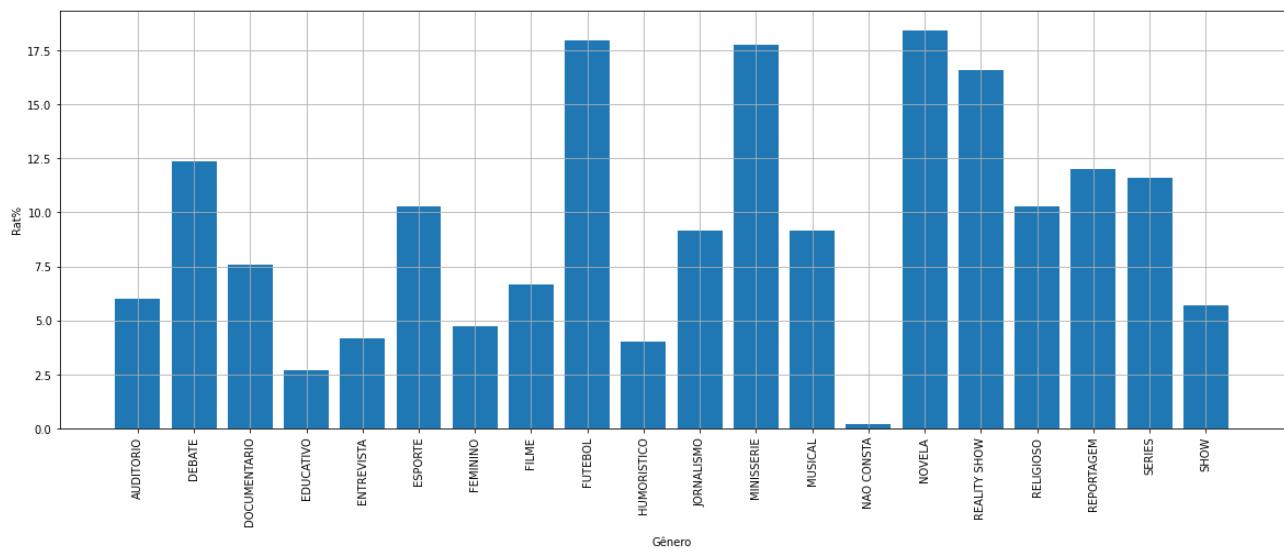


**Domingo**



A análise desses três gráficos mostra que, em geral, mulheres predominam na audiência geral por cerca de 1.5 Rat% durante todo o dia e quase 4 pontos durante o horário de pico. Esse resultado é surpreendente, dado que hoje apenas 7% das mulheres se declaram donas de casa e têm acesso a uma televisão o dia inteiro (“Parcela da população que se declara dona de casa cai para 7% em 26 anos”, 2019). Outra hipótese não validada é um maior interesse feminino por telenovelas. A exceção é o dia de domingo, quando homens ultrapassam as mulheres no horário de jogos de futebol.

### Audiência da Emissora 0 por gênero



Geramos esse gráfico com o objetivo de entender quais gêneros eram mais assistidos. Nesse sentido, definimos uma lista de “preferidos” pelo público com base nos picos relativos, chegando aos itens de Debate, Futebol, Minissérie, Novela, Reality Show, Reportagem e Séries. Curiosamente, o gênero Jornalismo, do qual esperávamos uma das maiores audiências, não figura nessas máximas relativas. Isso é tão inusitado para nós que suspeitamos que talvez o PeopleMeter não esteja classificando todos os blocos jornalísticos corretamente, dado que são eles que atraem as maiores taxas de audiência nas análises por hora.

### Considerações sobre o resultado desejado

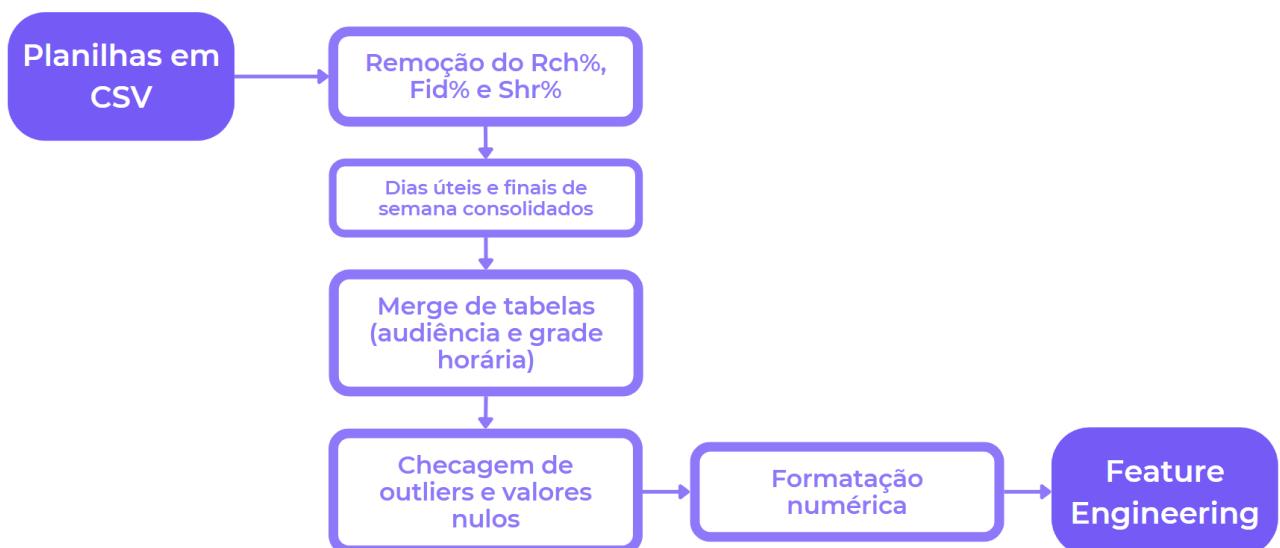
O resultado desejado da predição (saída) tem dois componentes. O primeiro e principal é o score de audiência esperado para um evento de determinadas características em certo dia e horário. Por ser um número dentre inúmeras opções, sua natureza é contínua (float). O segundo output planejado é uma lista das variáveis consideradas para chegar ao resultado principal e o peso de cada uma delas. Até o momento, não ficou claro se esse resultado será classificatório, partindo do princípio de que nem todas as variáveis possíveis serão utilizadas em cada predição, ou contínuo, caso os pesos de cada variável possam mudar de predição em

predição. Será necessário estudar mais sobre modelos preditivos e fazer as primeiras implementações para sanar essa dúvida.

## 4.3. Preparação dos Dados

### Pré-processamento dos dados

Os passos descritos abaixo referem-se às manipulações efetuadas para explorar dados, gerar gráficos e decidir quais features utilizar em nosso projeto. Nesse sentido, o objetivo era diminuir o tempo de processamento das planilhas no código, eliminar informações desnecessárias e formatar certos dados para otimizar filtros e diferentes visualizações antes de selecionar features, conforme demonstrado no fluxograma abaixo.



O código associado a cada passo pode ser conferido em nosso notebook do [Colab](#) e as planilhas resultantes de cada etapa estão disponíveis no Drive do grupo Jupyter.

### Anonimização dos dados

Dado que os dados fornecidos pelo Kantar IBOPE são confidenciais e exigem limitações contratuais com a TV Gazeta, principalmente no que tange às associações entre valores de audiência e emissora, realizamos um processo de anonimização sobre os arquivos antes de adicioná-los ao Google Drive e ao Colab. Este processo teve duas etapas:

1. No arquivo "TV\_Histórico.xlsx", substituímos todas as menções a emissoras por codinomes seguindo o padrão "Emissora A", "Emissora B" e "Emissora C";
2. No arquivo "grade\_Diária\_06\_2020\_a\_06\_2022.xlsx", não só realizamos as mesmas substituições de nomes de emissoras como também transformamos os títulos dos programas em "PROGRAMA 1", "PROGRAMA 2", "PROGRAMA 3", etc. Deve-se mencionar que esse método repete codinomes quando o programa em si se repete (por exemplo,

um evento de jornalismo diário denominado "PROGRAMA 56" será substituído por esse codinome todos os dias na planilha) e que a numeração reinicia a cada emissora (ou seja, "PROGRAMA 56" corresponde a programas diferentes em diferentes emissoras). Por fim, o arquivo original trazia programas e gêneros concatenados em uma única coluna, no formato "PROGRAMA / CATEGORIA". Porém, para facilitar a anonimização dos programas, segregamos, desde o início, essas informações em duas colunas para cada emissora: "EMISSORA X (Programa)" e "EMISSORA X (Categoria)".

OBS: Será enviado um dicionário com os dados e novas nomenclaturas ao parceiro.

### Otimização de processamento dos arquivos

O documento "audiencia\_original.csv", continha cerca de um milhão de linhas quando somadas todas as suas abas. Logo, o carregamento desse arquivo no código demandava muito tempo e recursos computacionais, além de gerar dataframes mais difíceis de manipular.

Sendo assim, nossa primeira ação em preparar os dados foi separar esse documento Excel em dezoito arquivos CSV, um para cada aba. Também eliminamos as colunas de audiência e demográficos que não se referissem ao "Rat%", métrica mais relevante. Desse modo, por enquanto, removemos os valores de "Shr%", "Fid%" e "Rch%", para ter acesso facilitado aos dados desejados e, também, diminuir o overhead do algoritmo através de um formato mais leve que o XLSX.

Contudo, durante o workshop com o parceiro da Sprint 2, aprendemos que também há interesse por parte da TV Gazeta em modelos que considerem os outros tipos de audiência. É algo que pretendemos explorar mais a fundo na próxima sprint.

### Formatação de datas

As datas no arquivo original estão formatadas como "dd/mm/aaaa", seguindo o padrão brasileiro. Inicialmente, consideramos desmembrar essa coluna em colunas de "Dia", "Mês" e "Ano"; entretanto, após alguns testes, percebemos que isso dificultava algumas análises mais profundadas dos dados no código. Por exemplo, se quiséssemos filtrar a primeira semana de cada mês, teríamos de aplicar diversas condições sobre as coluna tanto de dias quanto de meses. Em vez disso, decidimos deixar a planilha como está e apenas converter os valores das colunas em objetos de data no Python, através do método "pd.to\_datetime", quando fosse necessário manipular as datas de alguma forma. Um detalhe importante é que a conversão do XLSX para CSV com encoding UTF-8 formatou todas as datas, automaticamente, para "aaaa-mm-dd". Isso não afeta a conversão para objetos de data mencionada anteriormente.

### Agregando emissoras

Para agregar informações semelhantes, facilitar a análise por emissora e favorecer a separação de sets de treinamento no futuro, juntamos os CSVs de dias úteis, sábados e domingos de cada emissora em um único CSV para cada uma delas. Em mais detalhes, dado que com a etapa de

otimização conseguimos três arquivos por emissora (“Seg a Sex”, “Sáb”, “Dom”), decidimos, nesta fase, juntar esses arquivos em um só para cada emissora, ordenado cronologicamente.

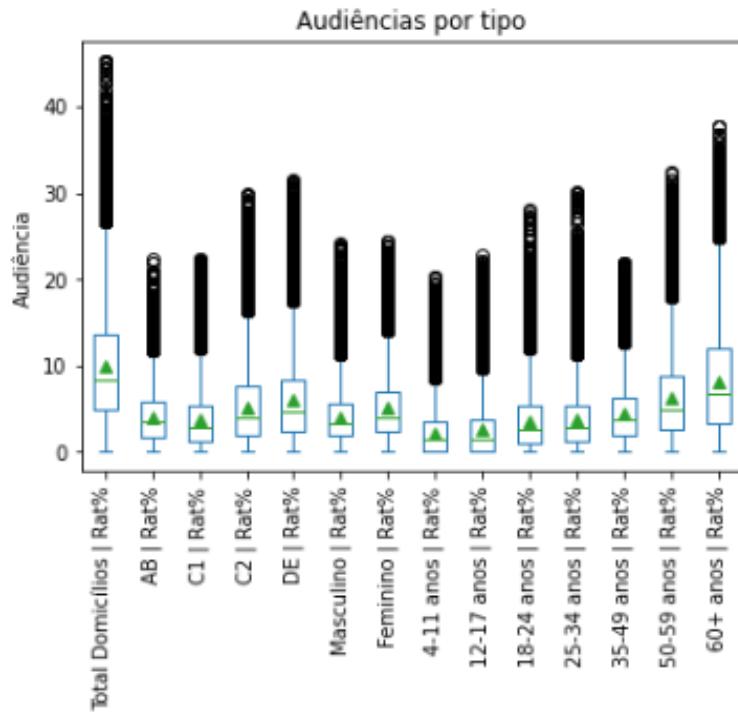
### Merge com grade horária

Com todos os dados de audiência por emissora compilados, restou mapear cada faixa horária de cada dia com o programa (e categoria) veiculados nesse período. Essas informações estão disponíveis no CSV “grade.csv”, que detalha o evento e o categoria para cada faixa horária e dia do período de 2020 a 2022. Sendo assim, agregamos uma coluna de “Programa” e uma de “Categoria” às planilhas das três emissoras contempladas no documento de grade horária, fazendo um match com as datas e horários presentes na origem e destino dessa transformação.

### Checando a existência de outliers

Para determinar se existem valores muito extremos nos dados que poderiam enviesar o modelo preditivo, tiramos a média e desvio-padrão de cada coluna em cada planilha. Então, conferimos se existia algum valor mais distante da média do que três desvios-padrão, pois este é o critério mais utilizado para identificar outliers. Para nossa surpresa, isso ocorreu com muita frequência, chegando a 21.000 ocorrências, o que nos causou estranheza. Parecia que estávamos perdendo dados demais nessa operação.

Para avaliar melhor o cenário, decidimos, então, gerar gráficos box-plot de todas as colunas. Esse tipo de gráfico é útil porque explicita não só a média de um dataset como também os percentis 25 e 75, a amplitude dos dados e os outliers de forma muito mais visual e intuitiva. Nesse modelo, outliers são valores além de  $1,5 * \text{IQR}$  (a amplitude entre os percentis 25 e 75) da média e geralmente aparecem como pontos isolados acima ou abaixo das caudas do gráfico, o que demonstra sua falta de encaixe com o resto do dataset.



Ao analisar os gráficos de nossos dados, entretanto, vemos que não há apenas alguns pontos isolados, e sim de uma incidência tão alta de valores extremos que mais parecem uma continuação das caudas do gráfico. Nesse sentido, é fácil visualizar que esses valores formam um padrão distinto no dataset e que eliminá-los descaracterizaria nossos dados. Afinal, outliers são, por definição, exceções. Aqui, é evidente que audiências altas estão longe de serem excepcionais.

Uma explicação contextualizada do porquê isso acontece é que os picos de audiência, apesar de regulares, acontecem com baixa frequência quando comparados com toda a amplitude de faixa horária. Em outras palavras, dentre as 24 horas possíveis de programação, apenas uma ou duas atingem esses picos vertiginosos. Assim, temos valores altíssimos e regulares, porém não frequentes o suficiente para puxar a média em sua direção, dado que as outras 22 horas do dia marcam scores muito menores. Logo, temos que a média permanece baixa, porém há uma significativa frequência de valores elevados além da amplitude esperada. Nesse contexto, decidimos manter todos os dados, pois não há outliers reais a serem removidos.

### **Checando valores nulos e ausentes**

A checagem de valores nulos, ausentes e/ou vazios foi feita através do método “`isnull()`”, da biblioteca Pandas. O resultado foi falso para as três planilhas de emissoras. Portanto, não foi necessário nenhum tratamento nesse quesito.

### **Checando categorias de baixa frequência**

Ainda na busca por outliers e exceções, percebemos que não deveríamos examinar apenas valores absolutos, mas também a frequência de certas variáveis categóricas. Nesse sentido, quisemos garantir que os gêneros de programa mencionados nas planilhas apareciam

com frequência suficiente para que conclusões pudessem ser tiradas de forma estatisticamente válida. Por isso, realizamos a contagem de cada categoria presente na grade horária e descobrimos que algumas delas possuíam frequências baixíssimas, com apenas 2 ou 7 ocorrências.

Assim, acabamos eliminando as categorias “NAO CONSTA”, “SORTEIO”, “OUTROS” e “TELE VENDAS”, pois não continham mais de 12 incidências no melhor dos casos. Nossa critério para essa decisão foi eliminar todas as categorias até a de “DEBATE”, que traz um aumento relativo de aparições, chegando a 36, e é inherentemente relevante às emissoras por suas ramificações políticas.

## Seleção de features

A finalidade da GIA é clara: prever o score de audiência para futuros programas com certas características (gênero) em certos dias e horários.

Nesse contexto, com a preparação anteriormente detalhada executada, pudemos começar a selecionar nossas features. Iniciamos pelas datas, pois seriam mais fáceis de manipular e preparar através das bibliotecas do Python. Em nossos estudos prévios, percebemos que o dia do mês não exercia influência significativa nas audiências (vide gráficos da seção anterior), contudo o dia da semana e o mês em si, sim. Assim, determinamos que essas duas informações (dia da semana e mês) deveriam ser contempladas nas features.

A fim de utilizar esses dados em um modelo preditivo, no entanto, precisávamos transformá-los em valores numéricos. Por isso, transformamos as variáveis categóricas associadas aos dias da semana (“Segunda”, “Terça”, etc.) em números de 1 a 7, seguindo a técnica de label encoding. Isso garantiu uma relação ordinal entre os valores, a qual desejávamos por conta da natureza ordinal dos dias da semana em si (segunda-feira está mais relacionada à terça-feira do que à sexta, por exemplo, e isso poderia trazer insights inesperados para o AI). Além disso, destrinchamos a coluna “Data” em uma de mês, composta por inteiros de 1 a 12, através da transformação das datas em objetos datetime no Python e o subsequente acesso ao atributo “mês” de cada um deles.

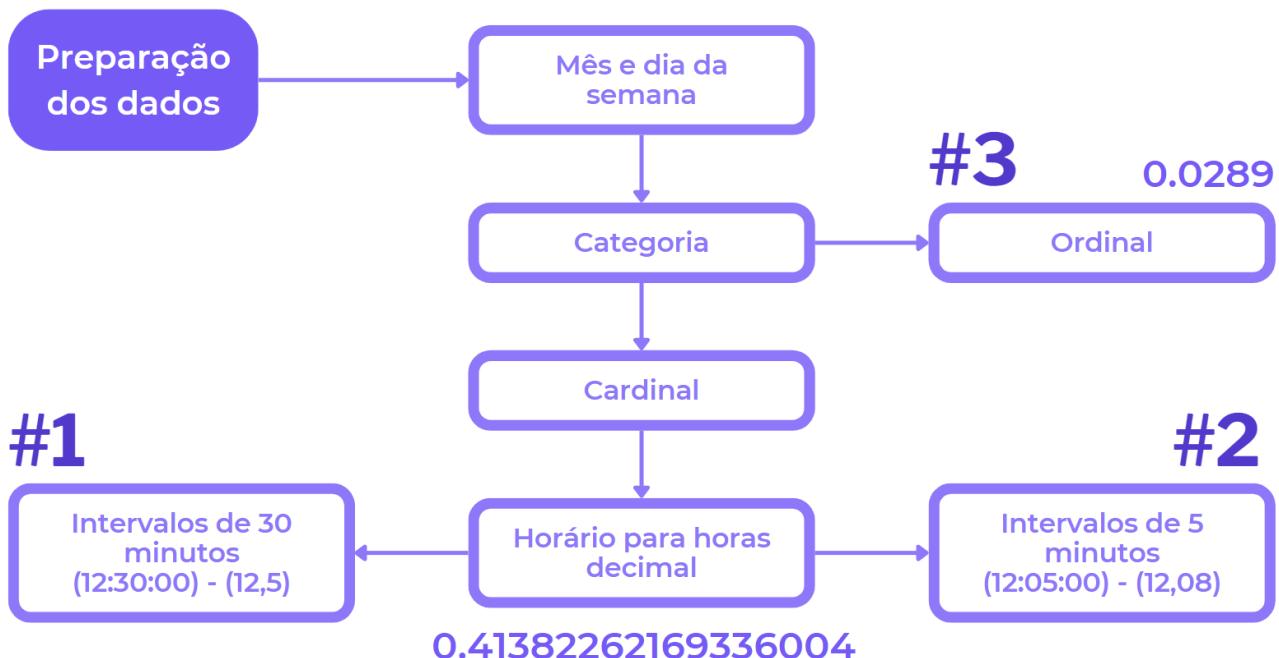
A próxima feature, por sua vez, deveria contemplar, de alguma forma, a faixa horária do programa. A seleção desse atributo causou muitas dúvidas e hipóteses. Pensamos em quebrá-lo em hora de início e hora de término, adicionar uma feature de duração, desmembrar o horário em um inteiro de hora e um inteiro de minuto, e muitas outras opções. Para fins de simplicidade e de ter um lugar pelo qual começar, todavia, resolvemos implementar dois tipos de entradas relacionadas a horário:

1. Horário em número decimal com intervalos de 30 minutos (12:30:00 se torna 12,5);
2. Horário em número decimal com intervalos de 5 minutos (12:05:00 se torna 12,08).

Não houve um raciocínio lógico muito extenso para a escolha desses formatos específicos além da noção de que precisávamos começar a testar alguma hipótese. Na teoria, todas as opções pareciam igualmente válidas. Portanto, escolhemos as mais simples para começar a medir e comparar desempenhos.

Por fim, para contemplar a categoria do programa, decidimos testar tanto a transformação por label encoding (ordinal, que dá um valor crescente começando em 1 a cada categoria) quanto a por one-hot encoding (cardinal, que cria colunas novas para cada variável única e categoriza os valores como 1 ou 0). Esperávamos que a versão cardinal tivesse resultados superiores desde o início, dado que categorias não possuem ordem inerente e essa premissa poderia confundir o AI. Ainda assim, a discrepância foi tão grande que acabamos testando essa abordagem apenas uma vez e desistimos logo em seguida.

Assim, ao final, idealizamos e testamos três hipóteses para feature engineering. A saída para todas elas consistia no Rat, a principal métrica analisada pelo parceiro e um bom ponto de partida para nossas explorações. Abaixo, detalhamos as hipóteses.



1. Entradas de dia da semana, mês, categoria cardinal e hora de início de meia em meia hora. Saída de Rat%. Predição com regressão linear;
2. Dia da semana, mês, categoria ordinal e hora de início de meia em meia hora. Saída de Rat%. Predição com regressão linear;
3. Dia da semana, mês, categoria cardinal e hora de início de cinco em cinco minutos. Saída de Rat%. Predição com regressão linear.

## Teste de hipóteses

A partir das features selecionadas, resolvemos testar cada hipótese para averiguar a eficácia das seleções em si e das transformações utilizadas (one-hot encoding/label encoding). Essa avaliação foi feita através de três métricas: erro médio quadrado, que penaliza mais erros maiores; erro médio, que mostra a discrepância média entre valor predito e valor real; e R<sup>2</sup>, que representa a eficácia da hipótese quando comparada a um modelo naive. Nessa última, objetiva-se chegar o mais próximo possível de 1.

Os resultados estão resumidos abaixo.

### Hipótese 1

As features da hipótese 1 foram dia da semana, mês, categoria cardinal e hora de início. Contudo, a hora de início teve uma redução de linhas de 30 em 30 minutos, tirando a média de cada período.

**Erro médio quadrado:** 32.804852251521474

**Erro médio:** 4.62607344055152

**R<sup>2</sup>:** 0.4184967210608739

Aqui, o erro médio é de quase 5 pontos de audiência. Para uma amplitude de cerca de 40 pontos, isso é um erro significativo, passando de 10%. Já o R<sup>2</sup> nem chega a 0,5, mostrando-se insuficiente.

### Hipótese 2

As features da hipótese 2 foram dia da semana, mês, categoria cardinal e hora de início. Contudo, a hora de início manteve a divisão de 5 em 5 minutos.

**Erro médio quadrado:** 24.80568954239913

**Erro médio:** 3.8310018190852384

**R<sup>2</sup>:** 0.4158580211459625

Para nossa surpresa, esse conjunto de features resultou em erros e R<sup>2</sup> similares à proposta anterior. Analisando o fato com mais atenção, percebemos que isso se dá porque a única diferença entre as hipóteses 1 e 2 é o fato de a primeira conter as médias da segunda a cada meia-hora. Logo, elas possuem valores proporcionais, cuja semelhança foi preservada no modelo preditivo. Apesar de o erro médio ter sofrido uma alteração maior, vemos que isso se dá apenas pela característica desse tipo de erro de penalizar mais erros maiores. Assim, percebemos que simplesmente mudar a periodização dos horários no futuro não altera significativamente o desempenho do modelo.

### Hipótese 3

As features da hipótese 3 foram hora de início (de 30m em 30m), dia da semana, mês e categoria ordinal.

**Erro médio quadrado:** 55.67440408917043

**Erro médio:** 6.238326034220553

**R<sup>2</sup>:** 0.028903147388637307

O resultado, apesar de conter erro médio similar, mostra-se evidentemente fracassado quando o R<sup>2</sup> é analisado – muito, muito longe de 1. Isso provavelmente se deu porque o processo de regressão detectou “padrões” não existentes entre os valores ordinais das

categorias, como uma maior semelhança entre categorias 1 e 2, mesmo que elas não tenham nada em comum na realidade. Esse teste nos mostrou, portanto, a importância de se utilizar one-hot encoding em variáveis não ordinais.

Em suma, nossos testes de hipótese foram vitais em entender melhor o que devemos priorizar em nossas features e modelos. Com base nisso, pensamos em mais opções de entradas que pretendemos explorar nas próximas sprints, como a adição de duração, hora de término e também a conglomeração de horários em períodos maiores, tais quais “MANHÃ 1”, “ALMOÇO”, etc.

## 4.4. Modelagem

O processo de modelagem foi longo e com diversas reviravoltas. Iniciamos com três hipóteses de features (detalhadas na seção 4.3), algumas ideias de algoritmos e vagas suposições sobre quais métricas utilizar.

Por fim, testamos mais de 28 modelos, selecionando quais algoritmos tendiam a performar melhor com nosso dataset e entendido profundamente como cada métrica funcionava e flutuava perante mudanças de feature ou parâmetro. Assim, esse processo todo se deu em três rodadas de experimentação, cada uma rendendo novos insights e direcionando (ou redirecionando) nosso planejamento.

A tabela abaixo resume essas etapas, que serão mais detalhadas no decorrer desta seção.

| #        | Features   | Algoritmos                            | Métricas   | Insights                                      |
|----------|--|---------------------------------------|--|---|
| Rodada 1 | Hora de início (15 em 15m), dia da semana, mês e categoria | Régressão linear                      | $R^2$ , $R^2$ ajustado, erro médio quadrado e erro médio absoluto para regressão linear e random forest; | Comparando $R^2$ e $R^2$ ajustado, percebemos |
| Rodada 2 | Hora de início (15 em 15m), dia da semana, mês e categoria | KNN, árvore decisória e random forest | $R^2$ , $R^2$ ajustado, erro médio quadrado e erro médio absoluto para regressão linear e random forest; |   |

## Rodada 1

Nossa primeira rodada teve como objetivo inicial testar as features escolhidas na Sprint 2 (**hora de início, dia da semana, mês e categoria cardinal**, tendo o “**Total Domicílios I Rat%**” da Rede Gazeta como saída) no algoritmo de **regressão linear**.

A análise de regressão linear é utilizada para prever o valor de uma variável (dependente) com base no valor de outra (independente). Desse modo, seguindo a lógica, o primeiro passo é descobrir o  $f(x)$  que nos devolve o valor aproximado de  $y$  correspondente ao  $x$  de entrada. Assim, utilizando as tabelas desenvolvidas na preparação dos dados (4.3), separamos em valores de treino ( $x_{treino}$  e  $y_{treino}$ ) para obter o  $f(x)$ . Posteriormente testamos o modelo, comparando o valor predito com o valor real ocultado. A explicabilidade deste tipo de modelo se dá através dos coeficientes encontrados.

Uma ressalva importante é que, seguindo recomendações de padronização do orientador de turma, decidimos periodizar a coluna de “Hora Início” de 15 em 15 minutos, em vez de utilizar as versões de 5 em 5 min ou 30 em 30 min da Sprint 2.

Já no que tange a **métricas**, começamos analisando, nessa ordem de importância,  **$R^2$ ,  $R^2$  ajustado, erro médio percentual, erro médio absoluto**. A justificativa para essa seleção se dá, porque R avalia o quanto certo modelo explica as variações de um dataset. Como buscamos um modelo que se assemelhe ao máximo com a realidade, precisamos justamente daquele que melhor explique e se adeque aos dados recebidos.

Porém, sabemos também que o  $R^2$  nem sempre é confiável. Por ser uma métrica associada à variância, quanto mais features são adicionadas, mais ele cresce, visto que um maior número de entradas resulta em maior variância mesmo que as adições não sejam relevantes. Por esse motivo, achamos importante sempre consultar o  $R^2$  ajustado, que retira esse viés e mostra a adequação real do modelo, independentemente do número de features.

Por fim, determinamos interessante também o erro médio absoluto, porque ele, estando na escala do output desejado, oferece uma perspectiva mais intuitiva do quanto bem o modelo performa. Assim, torna-se possível definir, com mais precisão, uma faixa de tolerância de erro. Nessa mesma linha, resolvemos dar peso considerável também ao erro médio percentual, obtido através da divisão do erro médio absoluto pela média de previsões; desse modo, entende-se o melhor que cada erro representa quando comparado à escala total. Não consideramos o erro médio quadrado porque, apesar de útil em muitos casos, não nos serviria tão bem quanto as outras métricas devido aos motivos explicitados acima; assim, para fins de simplicidade, dado que já tínhamos quatro métricas elencadas, decidimos ignorá-lo.

Quanto a faixas de tolerância ou objetivos para essas métricas, nessa primeira rodada ainda não as tínhamos. Como não havíamos testado quase nada, não tínhamos expectativas embasadas do que poderíamos alcançar em cada etapa. Portanto, estávamos, em outras palavras, apenas experimentando diferentes modelos para descobrir o que era possível.

Nesse contexto, começamos aplicando o método de regressão linear da biblioteca Scikit Learn nas features originais, isto é, sem nenhuma modificação de escala. Os resultados foram os seguintes:

```
Erro médio quadrado: 38.84954918657377
Erro médio: 5.063221162450242
Erro médio percentual: 35.39%
R2: 0.4722920738828278
R2 ajustado: 0.4722920738828278
```

Baseando-nos nos resultados preliminares da Sprint 2, já esperávamos que a regressão linear não passasse de 0,5 para valores “crus”. Com tão poucas features, também não causou surpresa que o R<sup>2</sup> tivesse o mesmo valor; sua importância viria, na verdade, através da comparação com outros modelos, conforme mais entradas fosse adicionadas. Ademais, mesmo sem metas objetivas para as métricas, sabíamos que 35% de erro é uma valor muito elevado.

Por isso, resolvemos tentar novamente, porém normalizando os dados (isto é, forçando-os para uma escala de 0 a 1) dessa vez. Essa técnica evita que discrepâncias de escala entre colunas prejudiquem as previsões. Assim, dado que nossos campos têm certa variação de escala (0 e 1 para categorias; 1 a aproximadamente 45 para audiência, por exemplo), supomos que a normalização poderia nos ajudar a melhorar nossas métricas. Eis os resultados:

```
Erro médio quadrado: 0.019085870501925662
Erro médio: 0.11222507194200755
Erro médio percentual: 35.39%
R2: 0.472292073882823
R2 ajustado: 0.472292073882823
```

Apesar de os erros terem caído bruscamente, isso só se deu porque eles também foram adaptados para a escala de 0 a 1. Na realidade, a adequação do modelo permaneceu a mesma, como pode ser visto no erro percentual e métricas de R<sup>2</sup>. Logo, não foi a diferença de escalas causou resultados negativos.

A próxima etapa, nesse sentido, foi tentar lidar com a variância individual de cada coluna através da padronização, que substitui valores pelos seus z-scores (número de desvios-padrão da média). A suposição era que essa performance era tão boa quanto, senão melhor do que a última tentativa, por ser apenas outra maneira de lidar com variância, tal qual a normalização. Fazendo isso, obtivemos:

```
Erro médio quadrado: 2.0475457273416322e+20
Erro médio: 248378030.81966096
Erro médio percentual: 100.0%
R2: -2.0044504114610543e+20
R2 ajustado: -2.0044504114610543e+20
```

Para nossa surpresa, essa abordagem foi tão ruim que resultou em um  $R^2$  negativo e um erro percentual de 100%! Indubitavelmente este foi nosso pior modelo de todo o projeto. Uma possível explicação para essas métricas é o fato de todas as colunas de categoria, antes binárias (0 ou 1), terem se transformado em decimais conforme seus z-scores. Assim, criou-se novamente uma ordinalidade que provavelmente distorceu as previsões.

Analizando esses resultados, chegamos à conclusão de que a regressão linear não era o melhor modelo para nossas features. Mesmo que alguma manipulação melhorasse os resultados, esse aumento dificilmente chegaria a um valor absoluto de  $R^2$  desejável. Assim, decidimos focar nossos esforços em outros modelos.

## Rodada 2

Nossa segunda rodada teve como objetivo testar as mesmas features (**hora de início, dia da semana, mês e categoria cardinal**, tendo o “**Total Domicílios I Rat%**” da Rede Gazeta como saída) no algoritmos de **KNN, árvore decisória e random forest**.

Inicialmente, havíamos entendido que o método “score” do KNN e árvore de decisão, o mais recomendado para avaliar esses algoritmos, retornasse a “acurácia” do modelo, isto é, o número de acertos sobre o número total de tentativas. Porém, através da comparação desse método com os resultados de  $R^2$  para cada modelo, descobrimos que ambos os métodos sempre retornavam o mesmo valor. Pesquisando mais a fundo, aprendemos que eles, de fato, calculam o mesmo  $R^2$ . Portanto, mantivemos, para esta rodada, as mesmas métricas de  **$R^2$**  (representado pelo método “score” em alguns modelos),  **$R^2$  ajustado, erro médio percentual e erro médio absoluto**. Ademais, novamente iniciamos os experimentos sem metas objetivas em mente, pois nos faltava ainda exploração para saber o que era possível esperar dos modelos.

Para o primeiro teste de modelo, desenvolvido a partir das 3 hipóteses, utilizamos a regressão linear. Como em todos os treinamentos e teste que serão expostos nesta seção, seguiremos as seguintes etapas:

1. Divisão do dataset em variáveis de treino e teste.
  - a. X será reservado para as features de entrada
  - b. Y será reservado para os valores de saída

2. Utilização de uma biblioteca para realização do treinamento .fit() e predição .predict()
3. Avaliação do modelo de acordo com a proximidade dos valores preditos, a partir do uso das features de teste ( $x_{\text{test}}$ ), com os valores realmente existentes na tabela ( $y_{\text{test}}$ ), anteriormente ocultados ao modelo. A proximidade dos valores preditos com os reais foram medidos a partir do erro médio absoluto (Mean Absolute Error), erro quadrático médio (Mean Squared Error) e o  $R^2$ .

## Regressão linear

Primeiramente, seguindo a lógica da regressão linear, precisamos descobrir o  $f(x)$  que nos devolve o valor aproximado de  $y$  correspondente ao  $x$  de entrada. Sendo assim, utilizando as tabelas desenvolvidas na preparação dos dados (4.3), separamos em valores de treino ( $x_{\text{treino}}$  e  $y_{\text{treino}}$ ) para obter  $f(x)$ . Posteriormente testamos o modelo, comparando o valor predito com o valor real ocultado.

Para  $Y$ , que será o nosso output, designamos a coluna target “Total Domicílios Rat%”. Já para o nosso input  $X$ , designamos as nossas features “Hora Início”, “Dia da Semana”, “Mês” e categorias.

No desenvolvimento e avaliação deste modelo, usando a biblioteca scikit learn, obtivemos entre as 3 hipóteses testadas, um  $R^2$  de 0.4184967210608739 e Erro médio quadrado igual a 32.804852251521474 após o teste do modelo. Sendo esse o melhor resultado obtido entre as 3 hipóteses, concluímos que a regressão linear é apenas 41,8% superior ao simples modelo de reta média dos valores de audiência. Como esperado, a regressão linear não é a melhor solução para as nossas variáveis, pois elas não seguem uma linearidade.

## KNN

Testamos também o modelo KNN (K-nearest neighbors) da biblioteca scikit learn e desta forma, atribuindo diferentes quantidades de vizinhos ( $n_{\text{neighbors}}$ ) comparativos obtivemos diferentes resultados. Para o modelo KNN é importante notar que devemos prioritariamente realizar a comparação com uma quantidade ímpar de vizinhos, isso porque a conclusão que é tomada pelo modelo considera os valores da maioria dos vizinhos.

Inicialmente, para 3  $n_{\text{neighbors}}$  a acurácia da amostra de teste foi de 0.7791908299629887. Progredindo com a quantidade de vizinhos para 15, alcançamos uma acurácia maior de teste de 0.8194354877728393, porém comprometendo o tempo de processamento.

## Regression Trees

Apesar de termos conseguido um resultado satisfatório com o KNN, decidimos testar outros modelos a fim de aprimorar as nossas previsões.

Atendendo a não linearidade dos valores da audiência, as árvores de regressão apresentaram-se um modelo viável. As árvores de regressão, diferentemente da regressão linear, são capazes de dividir o dataset em intervalos de valores. Para cada novo intervalo é

criado um galho. Os resultados da predição serão as folhas das árvores, nós que não têm sucessor, enquanto que a raiz da árvore, antecessora de todos os galhos, será decidida de acordo com a variável do conjunto de dados com menor soma dos quadrados dos resíduos.

Para esse modelo também utilizamos a biblioteca scikit learn, obtendo uma acurácia no primeiro teste de 0.8328103278599521 sem limites de branches, e acurácia de 0.6245650395377198 com quantidade máxima de galhos (max\_depth) = 3.

As árvores de regressão, no entanto, sofrem de um forte problema. Por não haver um limite ideal estabelecido de branches de intervalo criados, esbarramos no possível problema de overfitting do conjunto de dados, como mostrado no primeiro caso testado, isto é, o nosso modelo de árvore é pouco adaptável a novos dados entrantes e super adaptado ao conjunto de treinamento.

### **Random Forest Regressor**

A Random Forest Regressor, como o próprio nome indica, se resume em múltiplas árvores regressivas. Para a random forest são geradas diferentes árvores de regressão, essas árvores são usadas para construir um modelo randômico que aproveita as melhores características de cada árvore. Consequentemente, evitamos também a ocasião de um possível overfitting causado por alguma árvore muito extensa.

Usando a biblioteca scikit learn, conseguimos uma precisão de 0.8327605651577861

### **LightGBM**

O LightGBM é um algoritmo baseado em histograma que coloca valores contínuos em compartimentos discretos, o que leva a um treinamento mais rápido e a um uso mais eficiente da memória. A estrutura usa um algoritmo de crescimento de árvore em folha, que é diferente de muitos outros algoritmos baseados em árvores que usam crescimento em profundidade. Os algoritmos de crescimento de árvores em folhas tendem a convergir mais rapidamente que os em profundidade. No entanto, eles tendem a ser mais propensos a sobreajuste.

Preencher as seções 4.4 e 4.5 do documento, descrevendo os experimentos realizados até o momento e as respectivas análises. Sua entrega deve descrever:

- quais os algoritmos que foram escolhidos como adequados ao problema e aos dados e por que? (seção 4.4)
- qual foi a estratégia de avaliação escolhida e por que? (seção 4.4)
- resultados preliminares obtidos (seção 4.4)

- análise dos resultados (Qual a qualidade dos resultados? Qual a taxa de erro? Os algoritmos escolhidos foram adequados? Forneça exemplos e justifique suas respostas) (seção 4.5)

Para a Sprint 3, você deve descrever aqui os experimentos realizados com os modelos (treinamentos e testes) até o momento. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus experimentos e resultados.

Para a Sprint 4, você deve realizar a descrição final dos experimentos realizados (treinamentos e testes), comparando modelos. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus experimentos e resultados.

## 4.5. Avaliação

Nesta seção, descreva a solução final de modelo preditivo, e justifique a escolha. Alinhe sua justificativa com a seção 4.1, resgatando o entendimento do negócio e explicando de que formas seu modelo atende os requisitos. Não deixe de usar equações, tabelas e gráficos de visualização de dados para melhor ilustrar seus argumentos.

- análise dos resultados (Qual a qualidade dos resultados? Qual a taxa de erro? Os algoritmos escolhidos foram adequados? Forneça exemplos e justifique suas respostas) (seção 4.5)

A seguir, a partir das rodadas descritas na seção 4.4, apresentaremos uma análise do panorama geral, com comparações entre diferentes modelos, determinação de quais obtiveram melhores resultados e conclusões finais.

Os métodos avaliativos para a utilização de diferentes modelos foram baseados na média dos erros ao quadrados (MSE), pelo R<sup>2</sup>

### 4.5.1. Regressão Linear

Começamos com o modelo de regressão linear, aplicando o método de Label-Encoding para tratar as strings referentes às categorias dos programas

→ Média dos erros ao quadrado

$$MSE = \frac{1}{n} \sum_{t=1}^n e_t^2$$

→ Resultado da Precisão

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

|              |           | PREDICTED CLASS |           |
|--------------|-----------|-----------------|-----------|
|              |           | Class=Yes       | Class>No  |
| ACTUAL CLASS | Class=Yes | a<br>(TP)       | b<br>(FN) |
|              | Class>No  | c<br>(FP)       | d<br>(TN) |

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+TN+FP}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F-measure} = \frac{2 \cdot TP}{2 \cdot TP + FN + FP}$$

$$\text{Cost} = TP \times Cost_{TP} + FN \times Cost_{FN} \\ + TN \times Cost_{TN} + FP \times Cost_{FP}$$

$$\text{Sensitivity} = \text{Recall}$$

$$\text{Specificity} = 1 - \frac{FP}{FP+TN} = \frac{TN}{TN+FP}$$

$$\text{False Positive Rate} = 1 - \text{Specificity}$$

## 4.6. Comparação de modelos

## 5. Conclusões e Recomendações

Escreva, de forma resumida, sobre os principais resultados do seu projeto e faça recomendações formais ao seu parceiro de negócios em relação ao uso desse modelo. Você pode aproveitar este espaço para comentar sobre possíveis materiais extras, como um manual de usuário mais detalhado na seção “Anexos”.

Não se esqueça também das pessoas que serão potencialmente afetadas pelas decisões do modelo preditivo, e elabore recomendações que ajudem seu parceiro a tratá-las de maneira estratégica e ética.

## 6. Referências

**State of Mobile 2022** - data.ai. Disponível em  
<<https://www.data.ai/en/go/state-of-mobile-2022/>>.

**AMANDA CHAN** 12 APRIL 2011. Discovery Reveals Why Old People Go to Bed Early. Disponível em:  
<<https://www.livescience.com/13666-older-people-sleep-wake-early.html>>.

**Parcela da população que se declara dona de casa cai para 7% em 26 anos.** Disponível em:  
<<https://www1.folha.uol.com.br/mercado/2019/08/parcela-da-populacao-que-se-declara-dona-de-casa-cai-para-7-em-26-anos.shtml>>. Acesso em: 14 ago. 2022.

## Anexos

Utilize esta seção para anexar materiais como manuais de usuário, documentos complementares que ficaram grandes e não couberam no corpo do texto etc.

## Jornada do Usuário

### Expectativas

- Fazer uma melhor seleção de programas;
- Identificar os fatores que levam um programa ser popular;
- Suprir demandas desconhecidas do público;
- Desenvolver dados argumentativos mais sólidos.



### Rodrigo, Gerente de Operação e Programação

**Cenário:** Rodrigo quer realizar um plano de ação mais objetivo para a audiência de um novo programa.

|  |  |
|--|--|
| <b>FASE 1<br/>(Organização de dados)</b> <p>1.Começa a organizar (quase que manualmente) os números de audiência, características dos telespectadores e horários que possam se relacionar com o novo programa;<br/>2. Para acessar esses dados (tabelas em Excel) precisa pedir acesso para outros setores.</p>          | <b>FASE 2<br/>(Análise)</b> <p>1.Confere mais cuidadosamente os dados selecionados, com viés de análise;<br/>2.Checa o numero de audiência, características do público e horário de antigos programas que se relacionam com o novo;<br/>3.Imagina o quanto bem esse novo programa poderia ser encaxado com base em experiências anteriores.</p>  |
| <b>FASE 3<br/>(Deduzindo)</b> <p>1.A partir do que se foi imaginado e concluído (tendendo ao objetivo), temos as primeiras propostas de encaixe da nova programação expositiva, a partir da audiência;<br/>2.Discute as novas propostas com o que foi formulado.</p>   | <b>FASE 4<br/>(Preparando)</b> <p>1.Organiza os dados coletados anteriormente para serem inseridos no modelo preditivo;<br/>2.Apresenta as circunstâncias e dúvidas trabalhadas para o modelo preditivo.</p>   |
| <b>FASE 5<br/>(Conferir e Estabelecer)</b> <p>1.Compara as suposições subjetivas já feitas com aquelas desenvolvidas no modelo preditivo, fazendo questionar o que se foi concebido anteriormente ou apenas reforçando;<br/>2.Com esse estudo mais completo, sugestões dessa nova grade de horário são apresentadas.</p> | <b>FASE 5<br/>(Conferir e Estabelecer)</b> <p>'Agora sim, embasado e completo. Estou satisfeito com o resultado do modelo preditivo!'</p> <p>'Inserir todos os dados no modelo preditivo é bem cansativo/chato'</p> <p>'Bom, contudo, seria ótimo ter argumentos mais sólidos para a implementação de um novo programa na grade. Ainda tenho minhas dúvidas em relação a essas novas propostas'</p> <p>'Pelo o que já se passou até hoje, como seria o desempenho de um novo programa nesse cenário? Imagino o quanto bem esse programa e nossa rede possa se tornar'</p> <p>'É essencial organizar todos os dados que possam se relacionar com o contexto de um novo programa, mesmo o processo sendo extremamente cansativo'</p> |

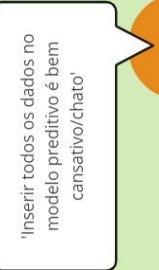
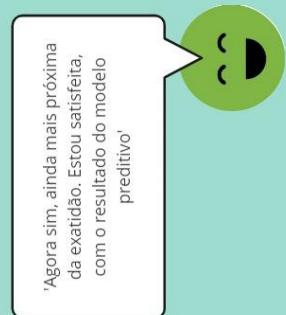
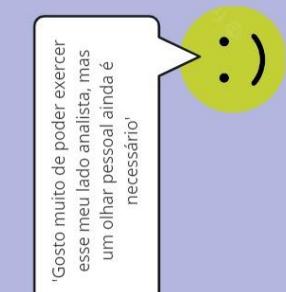
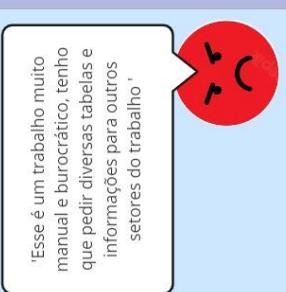
### Oportunidades

- Automatizar e deixar mais rápida a organização inicial de dados;
- Deixar mais prática a inserção de dados no modelo preditivo.

**Giovanna Mattos, Gerente Geral de Marketing**  
**Cenário:** Giovanna quer identificar exatamente (ou o mais próximo disso) o melhor horário para divulgar comerciais e propagandas dos novos programas.

### Expectativas

- Identificar fatores que mais chamam a atenção dos telespectadores em uma propaganda;
- Melhores evidências para alocação de propagandas.

|  |  |   |   |  |
|--|--|---|---|--|
| <p><b>FASE 1<br/>(Destrichando/Organizando)</b></p> <p>1.Após receber as informações do novo programa, confere quais programações já estabelecidas mais se assemelham a ele;<br/>2.Pede acesso aos dados (tabelas Excel) desses programas semelhantes (horário, público atingido, antiga forma de divulgação) para outros setores.</p> | <p><b>FASE 2<br/>(Análise)</b></p> <p>1.Confere mais cuidadosamente os dados selecionados, com viés de análise;<br/>2.Checa em quais programas e horários o público alvo está mais ativo;<br/>3.Compara como antigas divulgações desse gênero de programa foram feitas e seus impactos, mantendo acertos e evitando erros.</p> | <p><b>FASE 3<br/>(Deduzindo)</b></p> <p>1.A partir do que se foi imaginado e concluído na análise, temos as primeiras propostas de encaixe da divulgação do novo programa;<br/>2.Discorre as novas propostas com o que foi formulado.</p> | <p><b>FASE 4<br/>(Preparando)</b></p> <p>1.Organiza os dados coletados anteriormente para serem inseridos no modelo preditivo;<br/>2.Apresenta as circunstâncias e dúvidas trabalhadas para o modelo preditivo.</p>         | <p><b>FASE 5<br/>(Conferir e Estabelecer)</b></p> <p>1.Compara as suposições subjetivas já feitas com aquelas desenvolvidas no modelo preditivo, fazendo questionar o que se foi concebido anteriormente ou apenas reforçando;<br/>2.Com esse estudo mais exato e completo, a forma de divulgação desse novo programa é apresentada.</p> |
|  |  |   |  <p>'Inserir todos os dados no modelo preditivo é bem cansativo/chato'</p>  |  <p>'Agora sim, ainda mais próxima da exatidão. Estou satisfeita, com o resultado do modelo preditivo'</p>   |
|  |  |   |  <p>'Bom, mas ainda é algo um tanto que subjetivo, queria ter ainda mais exatidão nessa minha proposta de marketing em divulgação'</p>  |  <p>'Gostei muito de poder exercer esse meu lado analista, mas um olhar pessoal ainda é necessário'</p>  |
|  |  |   |  <p>'Esse é um trabalho muito manual e burocrático, tenho que pedir diversas tabelas e informações para outros setores do trabalho'</p> |  <p>'Agora sim, ainda mais próxima da exatidão. Estou satisfeita, com o resultado do modelo preditivo'</p>   |

### Oportunidades

- Deixar mais prático e evitar intermediários na organização inicial de dados;
- Tornar a inserção de dados no modelo preditivo menos manual e mais rápida.