hefEStos
Rede Gazeta

inteli
instituto
de tecnologia
e liderança

Controle do Documento

Histórico de revisões

Data	Autor	Versão	Resumo da atividade
08/08/2022	Marcos, Priscila, Maria Luísa, Matheus e Henrique	1.1	Criação do documento Edição do tópico 4.1.1 Edição do tópico 4.1.4 Edição do tópico 4.1.5
09/08/2022	Pedro Priscila Marcos	1.2	Edição do tópico 4.1.2 Edição do tópico 4.1.3 Inserção de dados nos subtópicos do tópico 4
10/08/2022	Maria Luísa Marcos Pedro	1.3	Revisão e conclusão dos tópicos do artefato 1 (4.1.1, 4.1.2, 4.1.3, 4.1.4 e 4.1.5)
15/08/2022	Maria Luisa Pedro Rafael Henrique Marcos Matheus	2.1	Fazer 4.2 - análise de dados
17/08/2022	Pedro Priscila Matheus	2.2	Persona Jornada de usuário
20/08/2022	Priscila	2.3	Formatação da documentação 4.1.2 - Descrição da matriz SWOT
24/08/2022	Maria Luisa	2.4	Revisão dos tópicos dos artefatos da sprint 1
25/08/2022	Maria Luisa Henrique Matheus	2.5	Formatação + passar para outro documento Realização do tópico 4.3

26/08/2022	Maria Luisa Priscila Falcão	2.6	Revisão de formatação, ortografia e conteúdo Tópicos 4.3.2 e 4.3.3

Sumário

1. Introdução

A **Rede Gazeta de Comunicações**, ou **Rede Gazeta**, é um conjunto de mídia brasileiro localizado no estado do Espírito Santo. Possuindo mais de 500 funcionários, a Rede Gazeta é o maior grupo de comunicação do Espírito Santo, a empresa foi fundada em 1928, com o jornal A Gazeta, porém apenas no ano de 1976 a TV Gazeta surgiu, aproveitando o grande crescimento dos meios de comunicação em massa. Atualmente, eles contam com a presença de 16 veículos de comunicação, abrangendo a TV, rádio e internet. Contudo, nos últimos anos, viu-se necessária a criação de um meio para melhorar a média de audiência dessa emissora

Nos últimos anos essa emissora cresceu bastante pelas suas produções próprias, porém com a grande competitividade das plataformas de streaming e das redes sociais, é de plena importância a constante pesquisa e análise de dados para suas próximas empreitadas. Por isso, o grupo hefEStos realizou a criação de um software que através da análise de dados e a partir da inteligência artificial realiza um modelo preditivo que ajudará na previsão da audiência de novos programas.

2. Objetivos e Justificativa

2.1. Objetivos

A Rede Gazeta procurou o Inteli para fazer esse projeto com os alunos, como o grupo hefEStos, para poder analisar e prever quais programas de TV são potenciais produtos para um investimento e expectativa maior.A empresa propôs a construção de um software que contará com machine learning para ter uma previsão de audiência para programas que já existem e que venham a ser lançados. O modelo preditivo que será entregue, deve receber alguns dados de entrada, sendo eles: data, horário e segmento do programa, com isso ele o software entregará uma previsão para o tipo de audiência, e tamanho da audiência medido em Rat%.

2.2. Justificativa

O grupo hefEStos propõe um modelo preditivo que entregará previsões se um programa específico terá ou não uma audiência adequada para tal(número de telespectadores), e também qual será o gênero, faixa etária e/ou a idade é o público alvo daquele programa piloto. Assim fazendo com que o cliente consiga otimizar seus gastos com programas, melhorar seu modelo preditivo e o modo como analisa seus dados .

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

3.2. Ferramentas

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

3.3. Principais técnicas empregadas

Descreva brevemente as principais técnicas empregadas, algoritmos e seus benefícios

4. Desenvolvimento e

4.1. Compreensão do Problema

4.1.1. Contexto da indústria

Com mais de 500 funcionários, a Rede Gazeta é o maior grupo de comunicação do Espírito Santo, a empresa foi fundada em 1928, com o jornal A Gazeta, porém apenas no ano de 1976 a TV Gazeta surgiu, aproveitando o grande crescimento dos meios de comunicação em massa. Atualmente, eles contam com a presença de 16 veículos de comunicação, abrangendo a TV, rádio e internet.

As principais concorrentes deste mercado de comunicação no estado do Espírito Santo são a TV Tribuna, afiliada do SBT, e a TV Vitória, afiliada da Record. Na grande maioria do tempo a TV Gazeta se apresenta mais forte que suas concorrentes, salvo algumas exceções ocasionadas por eventos específicos.

4.1.2. 5 forças de Porter

Ameaça de produtos substitutos - apesar de ser quase impossível a substituição da TV, por já ser um meio consolidado há várias décadas, é observada a transferência de certa parcela do seu público para os meios digitais.

Ameaça de entrada de novos concorrentes - o único tipo de ameaça sofrido pela TV são os meios digitais, que contam com conteúdos “on demand”, ou seja, que o usuário consegue escolher o que vai assistir. Porém, agora comparando a RedeGazeta com as outras emissoras, sua vantagem se estabelece pela confiança dos telespectadores, sendo ressaltada através dos dados.

Poder de negociação dos clientes - seus clientes, como os patrocinadores, antigamente não possuíam um grande poder de barganha, por não haver nenhum canal de comunicação tão abrangente quanto a TV, porém, com a evolução dos meios digitais, essa competição ficou mais acirrada, com outros locais para veiculação de anúncios publicitários, aumentando o poder de barganha a esses anunciantes, e assim, forçando aos detentores do meio a baixarem o preço.

Poder de barganha dos fornecedores- os principais fornecedores de uma emissora, são as empresas que trabalham com equipamentos de áudio e vídeo, por exemplo. Já que existem poucas emissoras no estado, que tem uma relevância grande no mercado, as empresas terceirizadas possuem baixo poder de barganha, uma vez que a demanda de consumo desses produtos não é tão alta. Dessa forma, há uma maior dependência dos fornecedores para com a emissora, do que o contrário.

Rivalidade entre os concorrentes - essa força se refere a necessidade de antecipar tendências e estar sempre atualizado com as novidades do mercado. A solução desenvolvida pelo grupo hefEStos atuará mais fortemente nesse aspecto de ficar a par das novidades do mercado, ajudando a reação do público na previsão do score de audiência de um novo programa.

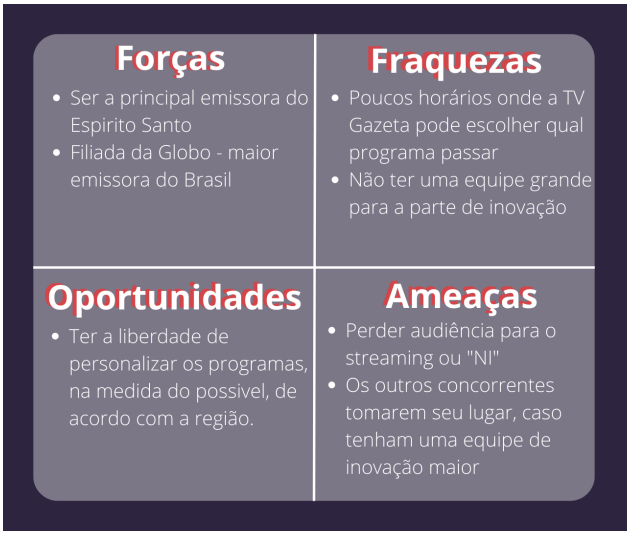
4.1.2. Análise SWOT

Com as análises feitas, foi possível verificar algumas forças, fraquezas, oportunidades e ameaças que o negócio apresenta. Dentro de *forças*, uma avaliação interna que mensura a capacidade da empresa de operar no mercado, visando a maior possibilidade de sucesso, observou-se, que com a aplicação da ferramenta desenvolvida, se terá um melhor uso e facilidade de análise dos dados que foram coletados, além de praticidade para análise caso queiram analisar outras grades e programas, atém mesmo de emissoras concorrentes.

Já em *fraquezas*, fatores que se apresentam como ponto de melhoria frente ao negócio, nota-se falhas no sistema de coleta de dados do IBOPE, uma vez que é feito em pequena escala, contando com a colaboração dos espectadores voluntários e correndo o risco de estarem viciados. Ademais, ainda pode ocorrer variação de dados pela subjetividade de quem está assistindo, para ilustrar, tome-se casos em que o segmento é bem recebido pelo público, mas ocorre apenas mudança de apresentador e sofre uma queda brusca.

Considerando fatores externos, tratando de *oportunidades*, constata-se a possibilidade de conseguir aumentar a audiência de programas já existentes, realocando os mesmos para horários mais adequados ou os direcionando para público específico, e ter uma previsão da repercussão do programa antes de ir ao ar. Ainda sobre elementos externos que podem afetar o negócio, tem-se as ameaças, que colocam a organização em posição frágil frente ao mercado, na qual foi possível diagnosticar sistemas de coletas e uso de dados mais eficientes de concorrentes como o YouTube e a Netflix.

Figura 1 - Matriz SWOT



Fonte: do próprio autor (2022).

4.1.3. Planejamento Geral da Solução

4.1.3.1. Dados disponíveis

Para o desenvolvimento da aplicação a Rede Gazeta disponibilizou alguns dados referentes à audiência, são esses:

- Datas;
- Hora de início;
- Emissoras;
- Dias da semana;
- Porcentagens utilizadas:
    - Rat%
    - Shr%
    - Rch%
    - Fid%
- Total de domicílios;
- Caracterização do público telespectador:
    - Classe:
        - AB;
        - C1;
        - C2;
        - DE;
    - Gênero:
        - Masculino;
        - Feminino;
    - Faixa etária:
        - 4-11 anos;
        - 12-17 anos;
        - 18-24 anos;
        - 25-34 anos;
        - 35-49 anos;
        - 50-59 anos;
        - 60+ anos.
- Grade de programação de cada emissora;

Divididos nas seguintes emissoras:

- Emissora 1;
- Emissora 2;
- Emissora 3;
- Canais pagos;
- Total Ligados Especial (TLE);
- Não identificado (NI);

Além disso, é dividido em dias da semana:

- Segunda a sexta;
- Sábado;
- Domingo.

4.1.3.2. Solução proposta

Sabendo da necessidade do parceiro por uma análise preditiva da audiência da Rede Gazeta, a solução proposta busca prover por meio de uma Inteligência Artificial, descrever a pontuação que determinado programa terá, quando lançado, de acordo com o horário, dia da semana, eixo e público especificado.

4.1.3.3. Tipo de tarefa

O tipo de método que será empregado será o de *Regressão,* pois iremos estimar dados de audiência de acordo com os valores de entrada que foram coletados anteriormente.

4.1.3.4. Como a solução deverá ser utilizada

O usuário deverá inserir os horários, a data, o segmento do novo programa e as características do público e a solução irá retornar uma previsão dos pontos de audiência, junto do peso de cada variável no modelo preditivo.

4.1.3.5. Quais são os benefícios trazidos

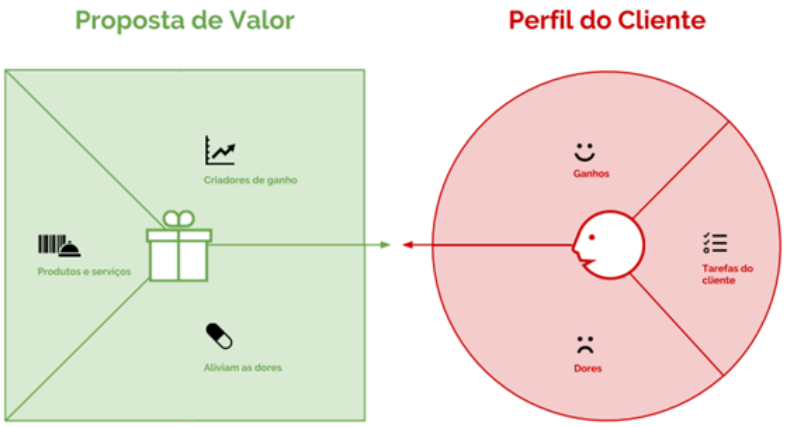
Os benefícios trazidos pela solução são os de melhorar a capacidade na realização da análise dos dados, de ajudar em previsões para lançamentos de programas futuros, além de promover uma melhora no desempenho de programas já existentes.

4.1.3.6. Qual será o critério de sucesso e qual será a medida utilizada

O critério utilizado como parâmetro de sucesso será a comparação com os valores fornecidos pelo banco de dados. Nesse, utilizamos parte do banco de dados para o treinamento do modelo, e o restante foi utilizado para comprovar o quão próximo foi o retorno do modelo com relação à realidade.

4.1.4. Value Proposition Canvas

Figura 2 - Value Proposition Canvas



Fonte: do próprio autor (2022).

4.1.4.1 Perfil do cliente

Tarefas do cliente:

- ➔ Criação de conteúdo para programas de TV;
- ➔ Analisar, de acordo com o histórico, se o programa iria se encaixar na programação.

Dores:

- ➔ Dificuldade em determinar o conteúdo adequado para agradar a audiência em um determinado horário;
- ➔ Alto investimento em programas que não repercutiram da forma esperada.

Ganhos do cliente:

- ➔ Melhora na acurácia da capacidade de prever a audiência de um determinado programa, baseado em algumas de suas informações como horário, data e tema. Adaptando, assim, o conteúdo, para aumentar o score de audiência.

4.1.4.2 Mapa de Valor

Produtos e serviços:

- ➔ Um modelo preditivo que recebe informações básicas de um possível novo programa,

e retorna um score de audiência e as principais variáveis que pesaram nesse.

Analgésicos/alívio das dores:

- ➔ Previsão acurada do possível sucesso de conteúdos, antes da produção.

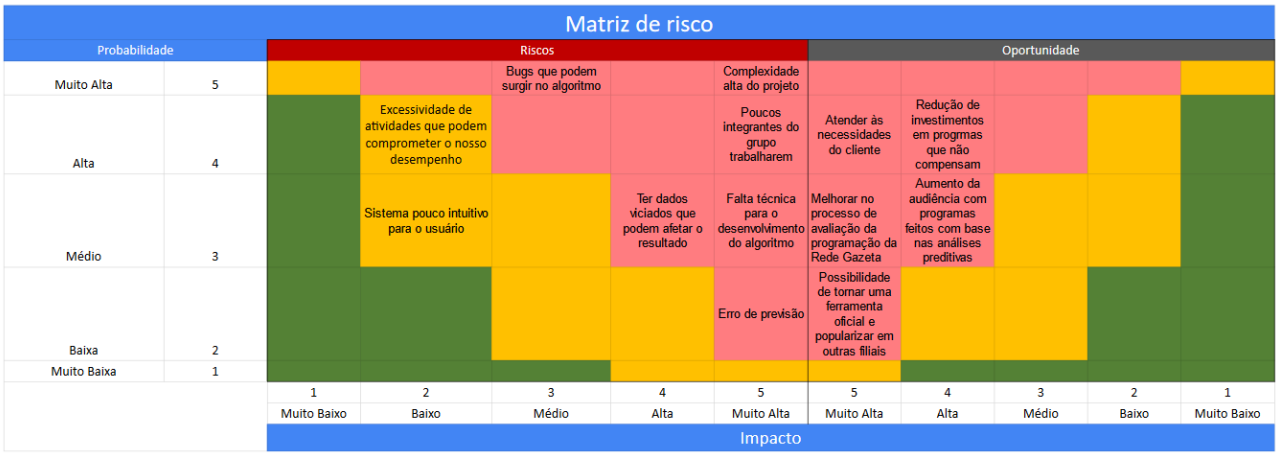
Criadores de ganhos:

➔ Evita gastos com programas de baixa audiência;

➔ Fornece informações de conteúdos capazes de maximizar a audiência.

4.1.5. Matriz de Riscos

Figura 3 - Matriz de risco.



Fonte: do próprio autor (2022).

4.1.6. Personas

Figura 4 - Persona

NOME: Ayumi Sato

IDADE: 30 anos

OCUPAÇÃO: Produtora de programa

Biografia:

- Se formou em Jornalismo
- Produz programas para uma filial
- Trabalha na mesma emissora a 10 anos

Características (personalidade, conhecimentos, interesses, habilidades):

- É apaixonada pelo seu estado natal
- Ama a tecnologia e acredita que pode ser usada em seu favor
- É uma das maiores produtoras de programa da empresa
- Está sempre disposta em ajudar

Motivações com a IA:

- Ter noção de quais estilos de programas o público prefere
- Ter diagnóstico de como será o desempenho daquele novo programa
- Vender a ideia para a emissora

Motivações com o problema:

- Poder melhorar suas ideias de acordo com a previsão
- Atingir públicos diferentes

Dores:

- Não ter uma noção de como o público irá reagir com o programa
- Não ter um sistema rápido e fácil para utilizar

miro

Fonte: do próprio (2022).

Figura 5 - Persona

NOME: Arthur Silva

IDADE: 35

OCUPAÇÃO: Gerente de Operação

Biografia:

- Engenheiro de computação
- Trabalha em uma filial de uma emissora
- Trabalha na mesma emissora a 7 anos

Características (personalidade, conhecimentos, interesses, habilidades):

- Um apaixonado por tecnologia
- Sempre amou consumir TV aberta
- Não se adaptou com Streaming
- Apaixonado por matemática e dados

Motivações com a IA:

- Conseguir prever quais programas devem ter um maior investimento
- Análise de desempenho
- Mostrar para os patrocinadores a previsão de impacto desse programa

Motivações com o problema:

- Poder investir mais em programas que podem ter um futuro melhor
- Atingir públicos diferentes
- Maiores patrocínios

Dores:

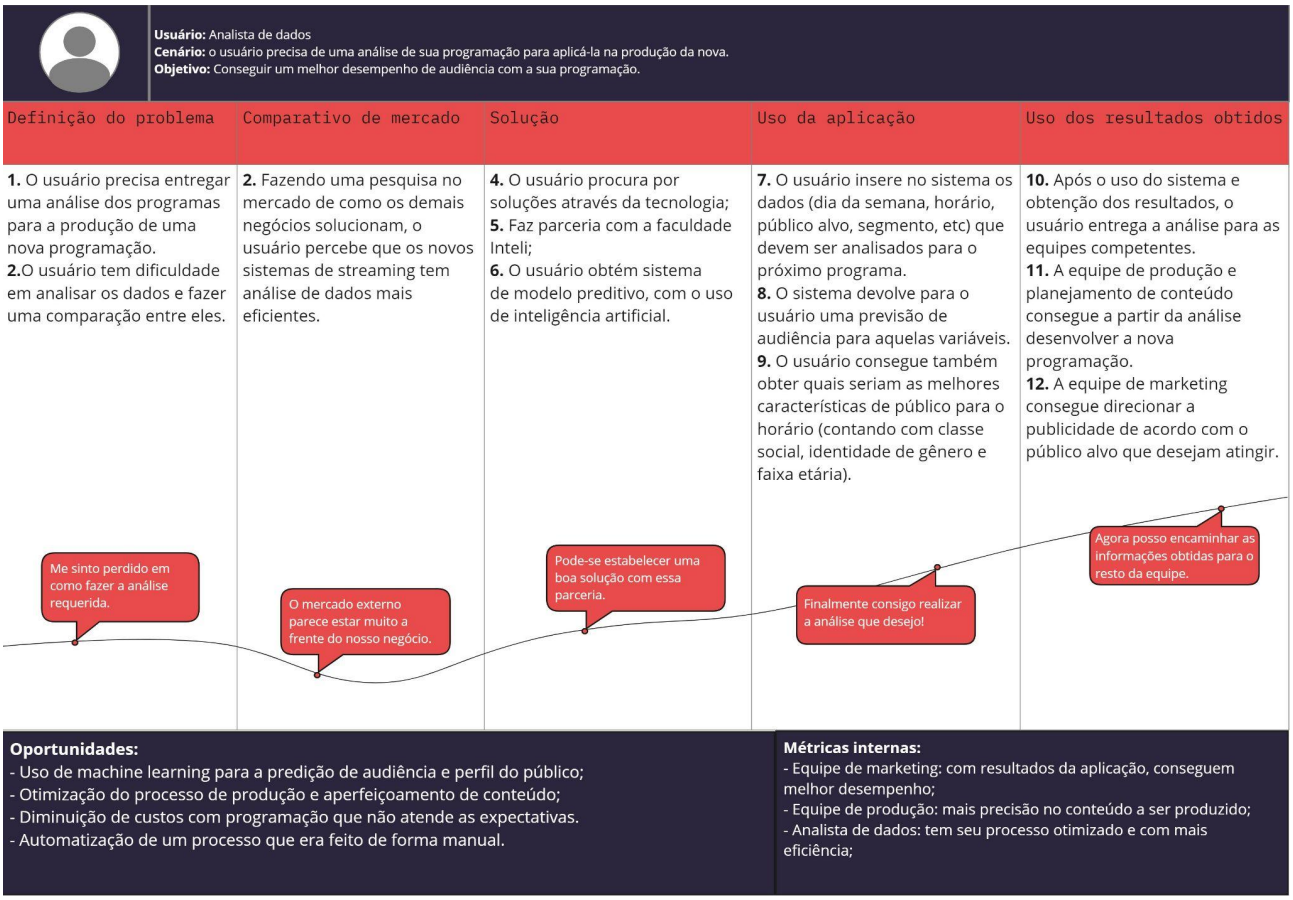
- Não ter uma noção de como o público irá reagir com o programa
- Dificuldade de analisar os dados e receber algum diagnóstico
- Não ter uma noção qual público será atingido

miro

Fonte: do próprio (2022).

4.1.7. Jornadas do Usuário

Figura 6 - Jornada do usuário.



Fonte: do próprio (2022).

4.2. Compreensão dos Dados

4.2.1. Descrição dos dados

Recebemos dados de diferentes emissoras no formato XLSX, que foi convertido posteriormente para CSV, baseados em pesquisas do IBOPE, utilizando parâmetros percentuais em: Rat%, essa medida é calculada a partir da quantidade de indivíduos/domicílios ligados em determinado evento de TV, sendo que 1 ponto de audiência se equivale a 1% do universo pesquisado. Esse é calculado da seguinte forma: Rat% = (Rat# / universo) x 100; Shr% descreve a participação da audiência em um determinado evento, sobre o total de televisores ligados, em um determinado período. Esse é calculado da seguinte maneira: Shr% = (Rat% / TLE%) x 100 ; Rch% é o total de indivíduos, ou domicílios, diferentes alcançados por pelo menos 1 minuto. Reforçando que o tempo total, nesse caso, não está sendo considerado, e sim o contato que houve com a programação, faixa horária, emissora, etc. Esse é calculado da seguinte forma: Rch% = número de telespectadores / universo x 100; Fid% ilustra a permanência dos telespectadores no evento em questão, ou seja, quanto tempo daquele programa foi consumido pelo público. Esse é calculado da seguinte maneira: Fid% = Rat% / Rch% x 100. Além disso, recebemos informações acerca do perfil da audiência, como gênero, faixa de idade e classe social.

4.2.1.1. Descrição de como os dados serão agregados/mesclados

Devido ao alto volume de dados, para melhor visualização, esses foram mesclados usando a média dos valores, e agregados a partir de uma funcionalidade da plataforma google sheets. Ademais, na segunda semana do projeto, depois de múltiplos pedidos da turma, foi adicionado uma nova planilha no nosso conjunto de dados, que contempla a grade horária, com as seguintes informações: praça, data, faixa horária e o nome do programa e segmento das três emissoras que estamos trabalhando(Emissora 1, 2 e 3). Com essas novas informações, vamos poder ter uma noção de qual programa está fazendo o maior sucesso, comparando com as outras emissoras, e assim analisar o porquê isso acontece, se é por conta do horário ou pela falta de concorrência, etc.

4.2.1.2. Descrição dos riscos e contigências relacionados a esses dados

Nesse contexto, durante a análise dos dados, foi verificado que há dois riscos a serem considerados na análise de tais dados: o primeiro risco é os dados tenham viés, já o segundo se trata da concorrência injusta na coleta dos dados.

No primeiro ponto é necessário ponderar que, no aparelho utilizado na medição da audiência, quando colocado no domicílio de cada família, é criado um perfil para cada integrante, com informações de classe social, gênero e idade. Quando a TV é ligada, quem está assistindo deve selecionar o seu próprio perfil, é nesse momento que pode acontecer o viés, já que uma criança, por exemplo, pode selecionar errado ou mais de uma pessoa pode estar vendo TV. Adicionalmente, a TV pode continuar ligada em um perfil originalmente correto, mas que já

não é válido, ou seja, outro integrante pode ter começado a ver TV.prestando atenção e não foi selecionado, ou até mesmo uma mãe/pai pode, na pressa, selecionar o perfil próprio para uma criança assistir.

O segundo risco é por conta da concorrência injusta, já que a emissora chamada de Não Identificado, abrange algo muito genérico, isto é, apesar de ser chamada de uma emissora, dentro dela ela é dividida por diversas plataformas de vídeos, tornando a comparação injusta para as emissoras 1,2 e 3, que estão separadas.

4.2.1.3. Descrição de como será selecionado o subconjunto para análise inicial

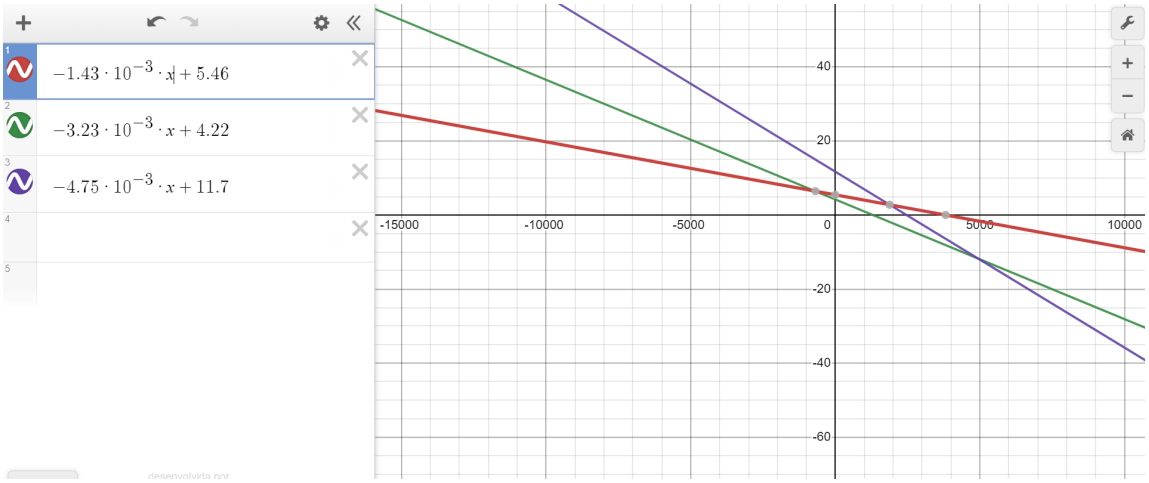
———————

4.2.1.4. Descrição das restrições de segurança

Para a elaboração do modelo, embora tenha sido concedido os dados de programação e audiência de algumas emissoras de TV aberta do Espírito Santo para a criação do modelo, foi proibida a publicação dos nomes das emissoras.

4.2.2. Descrição estatística básica dos dados

Figura 7 - Gráfico de tendência



Fonte: do próprio (2022).

- ➔ Emissora 1 = azul
- ➔ Emissora 2 = verde
- ➔ Emissora 3 = vermelho

Este gráfico ilustra a tendência de audiência nas emissoras em um intervalo de 2 anos, pode ser observado uma queda forte de audiência em todas, se destacando a maior queda na emissora 1.

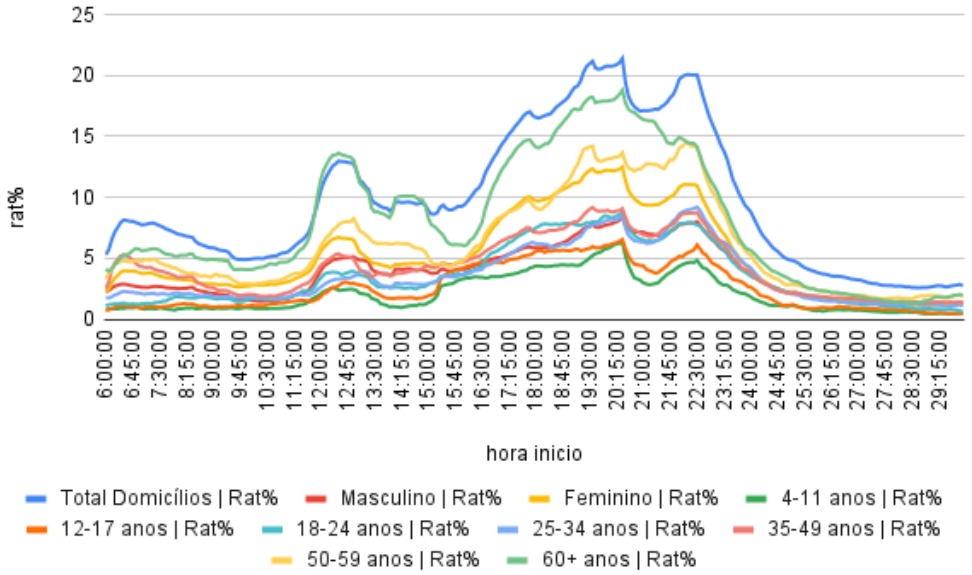
4.2.3. Dicionário dos dados

Parâmetros	Descrição	Como calcular
Rat%	Quantidade de indivíduos/domicílios ligados na TV (1 ponto = 1%)	Rat% = (Rat# / universo) x 100
Shr%	Participação da audiência em um evento, sobre o TLE de um período	Shr% = (Rat% / TLE%) x 100
Rch%	Total de domicílios ou indivíduos alcançados por 1 minuto ou mais	Rch% = número de telespectadores / universo x 100
Fid%	Permanência dos telespectadores naquele evento	Fid% = Rat% / Rch% x 100

4.2.4. Emissora 1

Segunda a sexta

Figura 8 - Audiência de segunda à sexta da emissora 1.



Fonte: do próprio autor (2022).

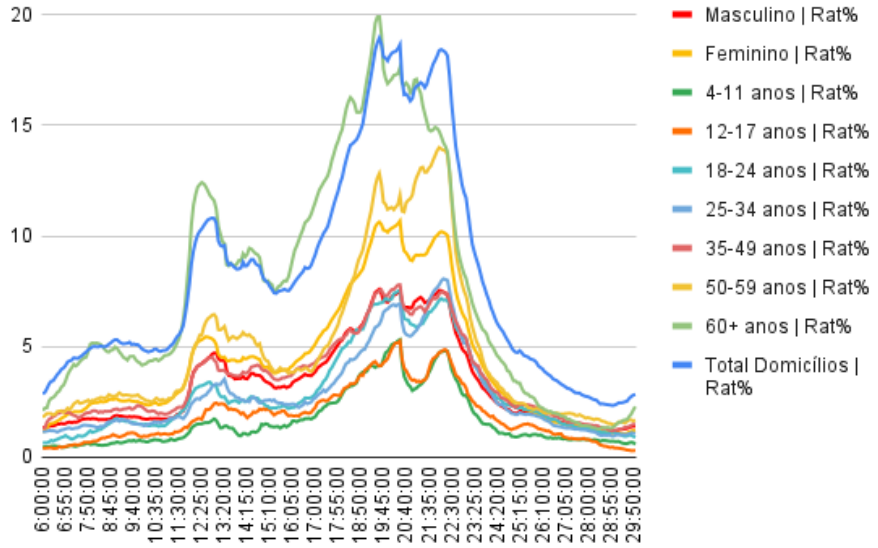
O primeiro gráfico que fizemos é da Emissora 1, juntando todos os dados de segunda a sexta, entre junho de 2020 e junho de 2022. Dividimos o Rat% em segmentos de idade, gênero e total de domicílios, e todos os dados são coletados a cada 5 minutos.

Analisando o gráfico, percebemos que em todos os horários o público feminino é predominante em comparação com o masculino. Seguindo a mesma ideia, o público com mais de 60 anos é predominante em comparação ao resto das idades, se mantendo estável durante toda a tarde, o único momento em que eles perdem a soberania é depois das 22:00, onde o público de 50-59 anos ganha.

Existem dois picos claros nesse gráfico: 1. No horário do almoço(aproximadamente às 12h), no momento em que está passando programa jornalístico regional. 2. Durante a noite (das 18h até aproximadamente 00h), que é o horário em que são exibidos programas de entretenimento como novelas e reality shows e jornais. Pode-se notar também que no horário de exibição de um jornal à noite (aproximadamente das 20h30 às 21h30), há uma redução do público.

Sábado

Figura 9 - Audiência de sábado da emissora 1.



Fonte: do próprio autor (2022)

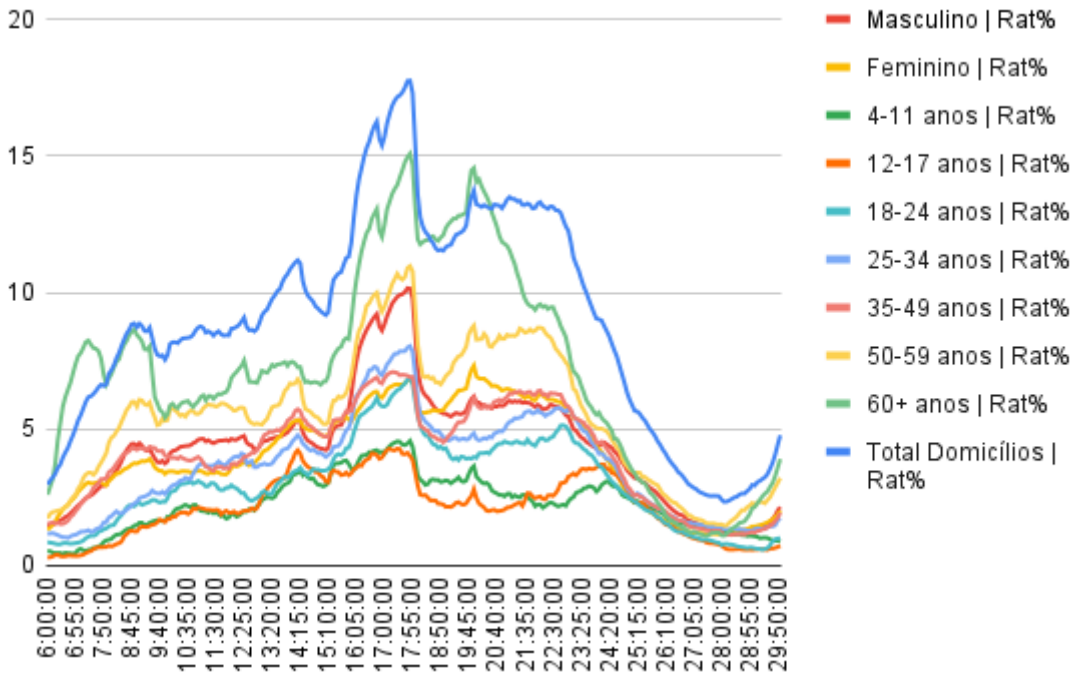
Fizemos um gráfico que reúne todos os sábados do mesmo espaço de tempo (junho de 2020 até junho de 2022), dividimos nos mesmo segmentos do gráfico anterior: todos os dados de Rat% com gêneros e idades e o total de domicílios.

Seguindo o mesmo padrão do gráfico anterior, as pessoas que têm mais de 60 anos continuam na soberania de audiência, só que percebemos que no sábado a diferença é bem menor comparado aos dias úteis. Além disso, em alguns horários essa faixa etária ganha do total de domicílios, como por exemplo entre 14:15 até 18:50.

As mulheres continuam com uma maior audiência do que homens, seguindo o mesmo padrão do gráfico anterior, porém em alguns horários, como 14:00 e na madrugada, essa diferença diminui bastante.

Domingo

Figura 10 - Audiência de domingo da emissora 1.



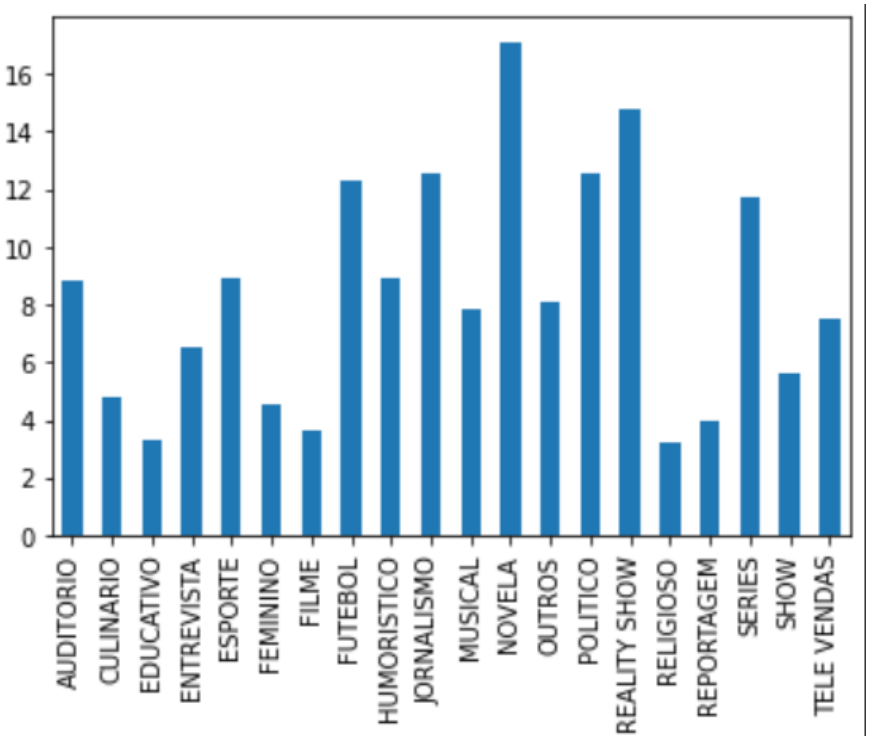
Fonte: do próprio autor (2022).

Fizemos um gráfico que reúne todos os domingos do mesmo espaço de tempo (junho de 2020 até junho de 2022), dividimos nos mesmo segmentos do gráfico anterior: todos os dados de Rat% com gêneros e idades e o total de domicílios.

Nesse gráfico, pode-se notar que a variação da audiência ocorre de outra forma, devido à diferença muito grande da grade de programação com relação aos dias úteis, sendo muito constante ao longo do dia com programações variadas de reality show, carros, rural,etc, exceto no horário do programa de futebol(das 15:50 às 18:05), em que há um pico em praticamente todos os nichos. Ainda nesse contexto, deve-se ponderar que logo após o término da programação de futebol há uma queda que leva a audiência de volta para o patamar anterior, e fica estável durante a noite com uma programação de auditório e de show.

Divisão por segmento

Figura 11 - Gráfico de audiência por segmento da emissora 1.



Fonte: do próprio autor (2022).

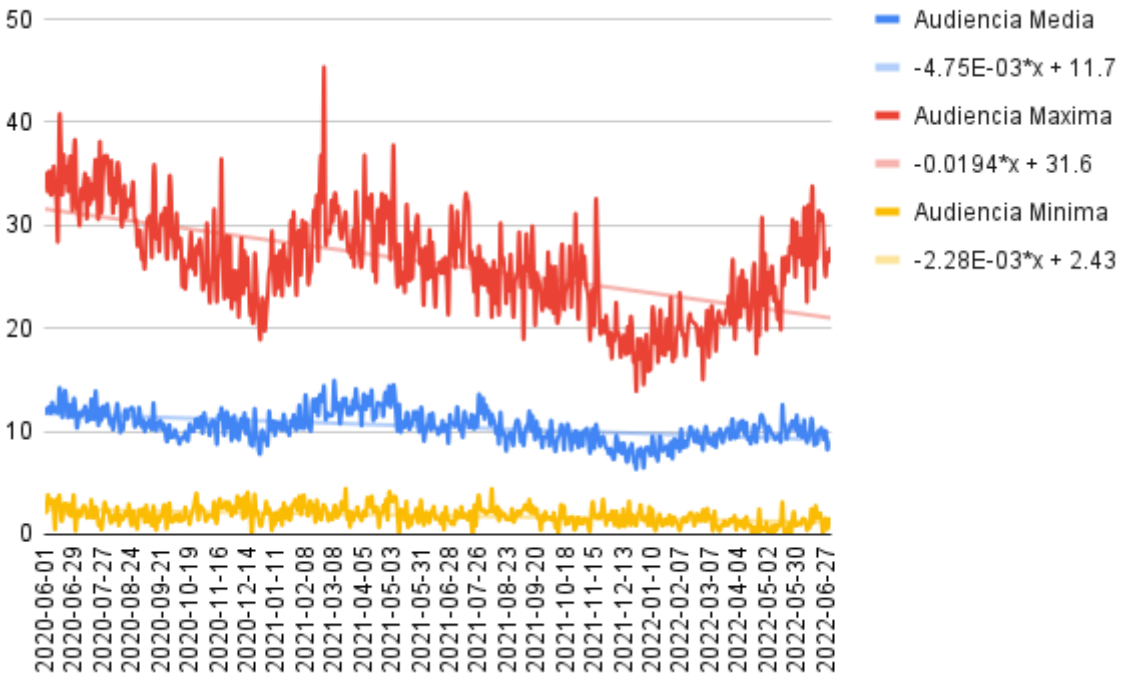
Quando recebemos os dados, cada programa estava separado em segmentos diversos, com isso fizemos um gráfico com o Rat% com o total de domicílios, assim os número no eixo y representam as porcentagens e o eixo x os segmentos. Esse gráfico representa somente os segmentos da emissora 1.

Podemos observar que os três segmentos com maior audiência são: novela, reality show, jornalismo e políticos, que estão empatados, e futebol, em ordem decrescente. Já os com pior audiência são: educativo, religioso, filme e reportagem, em ordem crescente.

Porém interpretando os dados podemos percebemos que essa alta de audiência no reality show pode acontecer por conta de um programa específico, que ocorre em alguns meses do ano, podem dar essa maximizada nos dados, sendo um ruído.

Segunda a Domingo - Audiência

Figura 12 - Gráfico de médias da audiência diária.



Fonte: do próprio autor (2022).

Esse gráfico nos mostra uma média de 3 dados importantes: audiência máxima, média e mínima. A primeira mede qual foi a maior quantidade de pessoas que estavam com a TV ligada na emissora 1 naquele dia específico, ou seja, o pico do dia. Já a segunda é a média de audiência naquele dia, que se mantém bem estável, comparando com o primeiro dado citado. O terceiro dado é exatamente o oposto do primeiro, então ele mede quanto que foi a menor quantidade de pessoas ligadas na TV naquele dia, ou seja, o ponto baixo do dia.

Podemos notar com clareza que, a audiência média e mínima, se mantém estável, quando comparado com a linha, que mostra a média daqueles valores. Já a audiência máxima isto não acontece, existem pontos específicos em que a audiência é visivelmente mais alta ou mais baixa. Notamos que esses picos mais altos são em momentos particulares, como final de algum reality show, final de campeonato de futebol, ou em feriados como o Natal.

4.2.4. Emissora 2

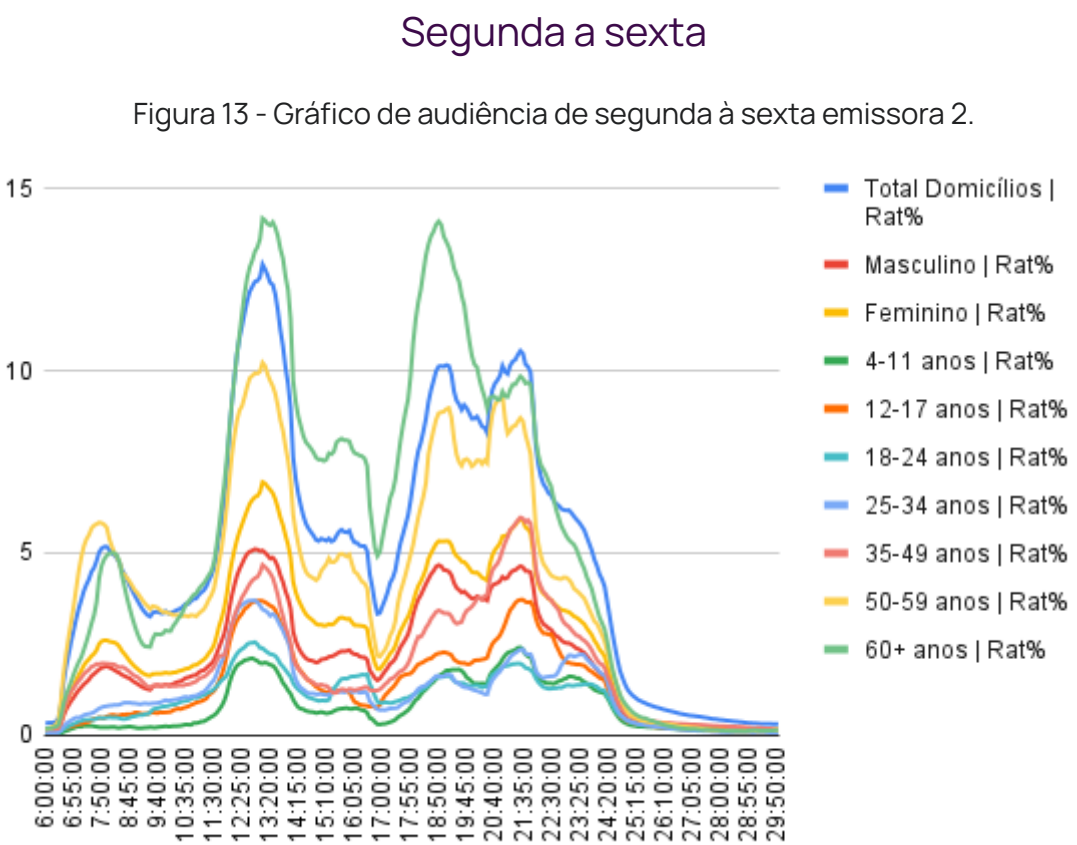


Figura 13 - Gráfico de audiência de segunda à sexta emissora 2.

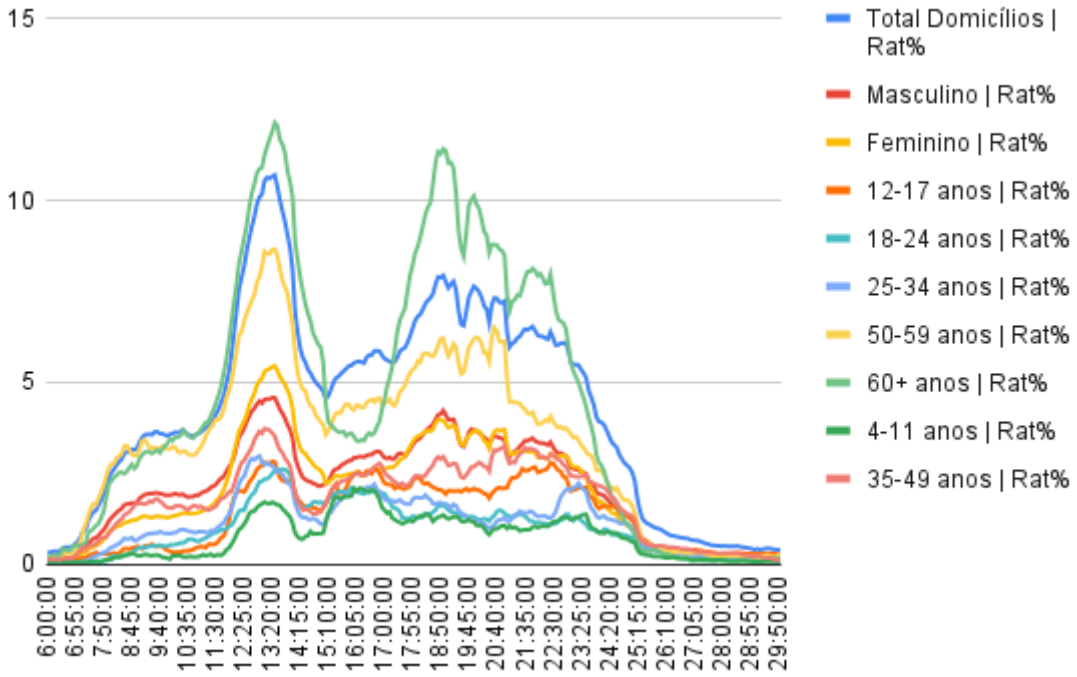
Fonte: do próprio autor (2022).

O gráfico acima representa a média de audiência média dos dias de segunda a sexta dos últimos dois anos na emissora 2. Sendo separado por idade, gênero e total de domicílios. Como a maioria dos gráficos que analisamos, esse se mantém no mesmo padrão: o sexo feminino ganha do sexo masculino em todos os horários, e o mesmo acontece com a idade 60+, porém algo se diferencia: as pessoas que tem 50-59 anos ganham de todas as outras idades, menos de 60+, e os dois gêneros e em alguns momentos até mesmo do total.

Os picos principais acontecem durante o horário de almoço (12:25 - 14:15) e no início da noite (18:50 - 20:40) e isso se dá em todas as linhas, em proporções diferentes. Por outro lado existem momentos que há uma queda muito grande, 14:15 - 15:10 e 17:00, isso pode ser por dois motivos: um deles é que as concorrentes podem estar passando algo que é mais interessante para esse público ou o programa que passa na emissora 2, nesses horários, não agrada o público.

Sábado

Figura 14 - Gráfico da audiência de sábado da emissora 2.

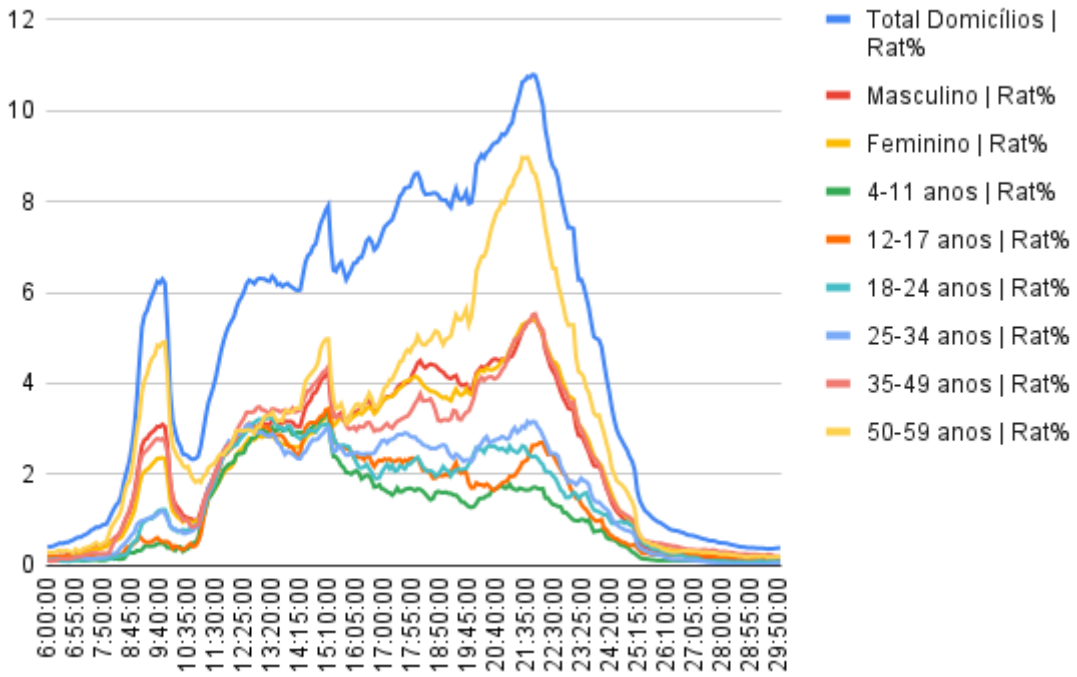


Fonte: do próprio autor (2022).

Este gráfico mostra a média de audiência por horário nos sábados da emissora 2 dividido por faixas etárias. Pode-se perceber que ao longo do dia inteiro a audiência predominante é de 50+ anos. Nota-se também dois picos de audiência, o primeiro às 13:20h e o segundo por volta das 19:00h, sendo o segundo mais dispersado. Os picos do gráfico são muito mais definidos nas faixas etárias mais velhas, ou seja, apesar de o público jovem ter menor audiência em todos os horários sua audiência é mais constante ao longo do dia.

Domingo

Figura 15 - Gráfico de audiência de domingo da emissora 2.



Fonte: do próprio autor (2022).

Juntando todos os dados de Domingo, entre junho de 2020 e junho de 2022. Dividimos o Rat% em segmentos de idade, gênero e total de domicílios, e todos os dados são coletados a cada 5 minutos.

Analisando o gráfico, percebemos que em todos os horários o público feminino é predominante em comparação com o masculino. Seguindo a mesma ideia, o público com mais de 60 anos é predominante em comparação ao resto das idades, se mantendo estável a partir do 12:00 durante toda a tarde e à noite, o único momento em que eles perdem a soberania é depois das 7:00, onde o público de 50-59 anos ganha.

A audiência é mantida alta das 12:00 até 23:00 em média, possuindo um pico às 13:00, que pode ser em decorrência da hora média do almoço. Sendo o único momento de baixa audiência durante a madrugada e de manhã até às 12:00.

4.2.5. Conteúdo Não Identificado

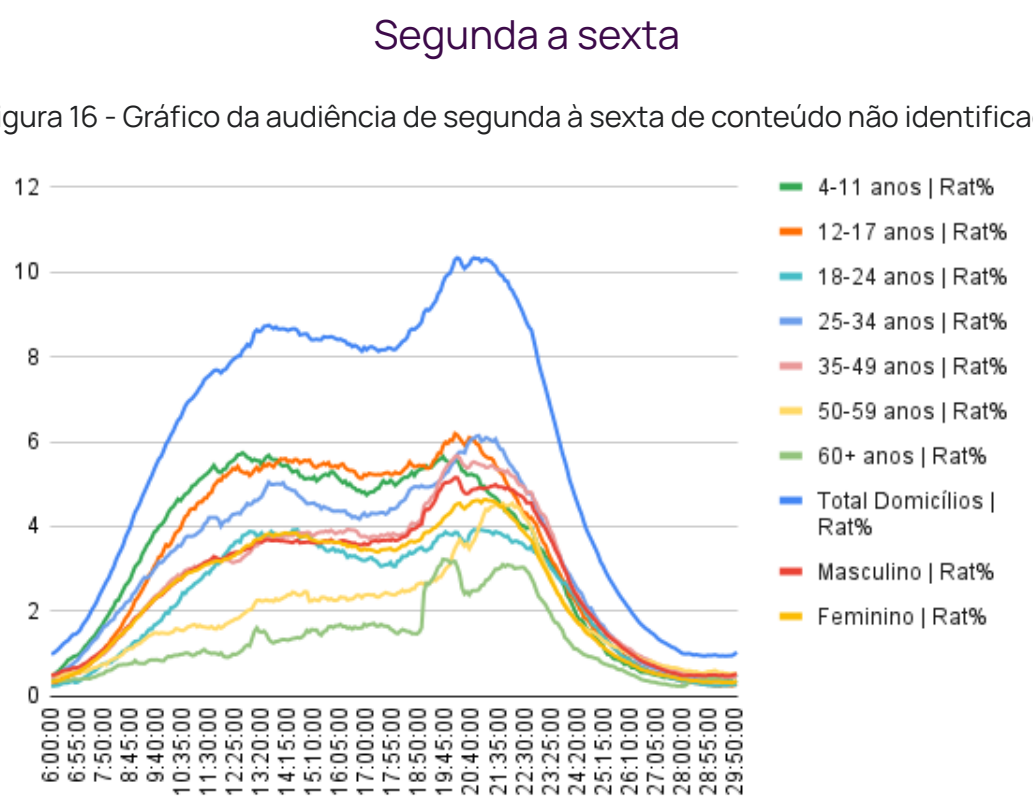


Figura 16 - Gráfico da audiência de segunda à sexta de conteúdo não identificado.

Fonte: do próprio autor (2022).

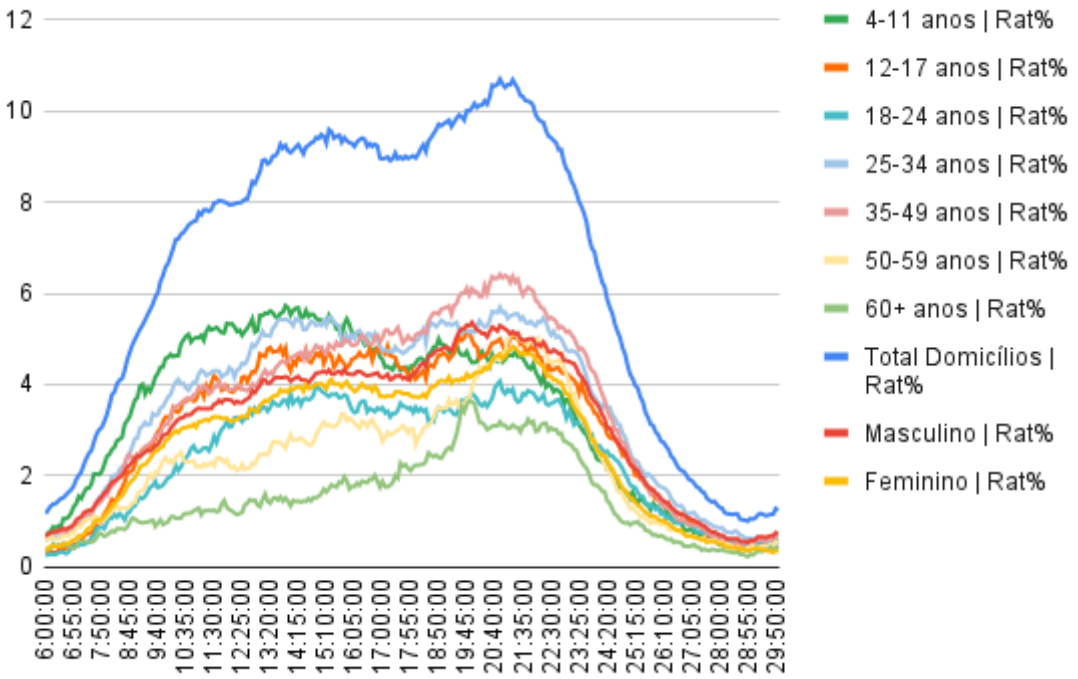
O primeiro gráfico que fizemos é a junção de todos os dados de segunda a sexta, entre junho de 2020 e junho de 2022. Dividimos o Rat% em segmentos de idade, gênero e total de domicílios, e todos os dados foram coletados com um intervalo de 5 minutos.

Analisando o gráfico, é possível perceber que a partir das 19:00 até 00:00 o público masculino é predominante, outro ponto que pode se observar é que o público entre 12-17 anos é predominante em todos os horários.

Existem dois picos claros nesse gráfico: 1. No final da tarde (por volta das 18:00) o único público que tem um pico menos relevante são as pessoas de 18-24 anos. 2. Durante a noite (por volta das 20:40)nos públicos com 50-59 anos e 60+ anos, também se vê um pico não muito relevante entre as pessoas com 18-24 anos.

Sábado

Figura 17 - Gráfico de audiência de sábado do conteúdo não identificado.



Fonte: do próprio autor (2022).

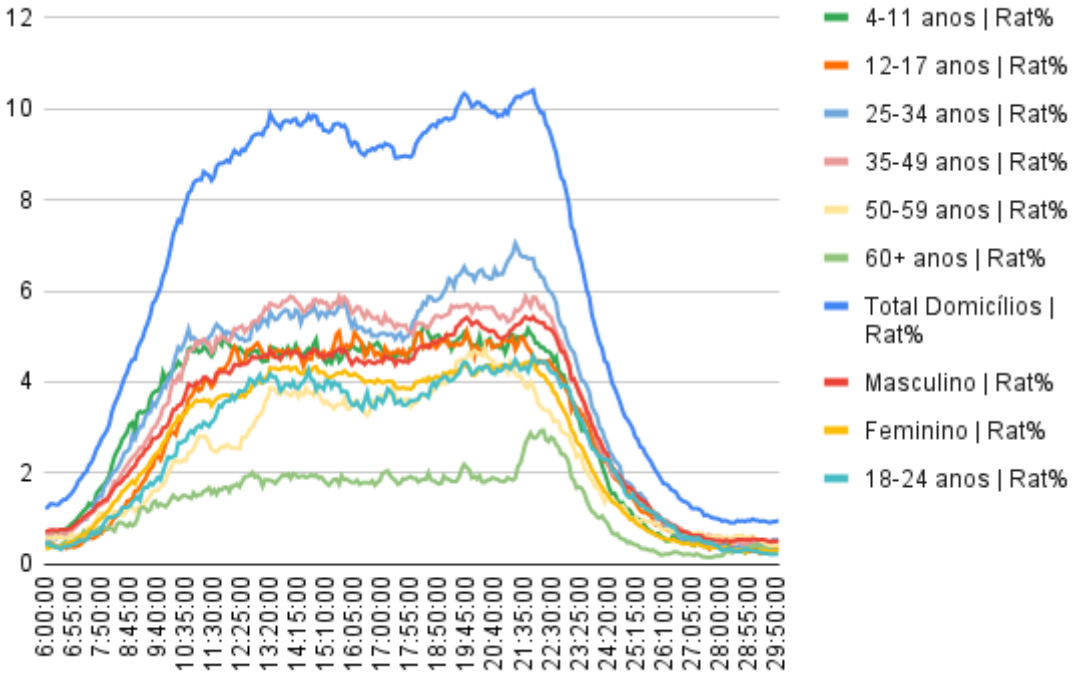
Fizemos um gráfico que reúne todos os sábados do mesmo espaço de tempo (junho de 2020 até junho de 2022), dividimos nos mesmo segmentos do gráfico anterior: todos os dados de Rat% com gêneros, idades e o total de domicílios.

Diferente do gráfico anterior, o público masculino no se mantém predominante durante o dia inteiro aos sábados, outro ponto é que o público de 4-11 anos se mantém predominante durante o período da manhã até a tarde(por volta de 16:40) que é quando o público de 35-39 anos vira o público predominante durante o resto do dia.

Também é possível perceber um aumento de audiência na maior parte dos públicos por volta de 19:45, menos no público infantil(4-11 anos) que tem um leve pico de audiência que depois começa a cair .

Domingo

Figura 18 - Gráfico de audiência de domingo do conteúdo não identificado.



Fonte: do próprio autor (2022).

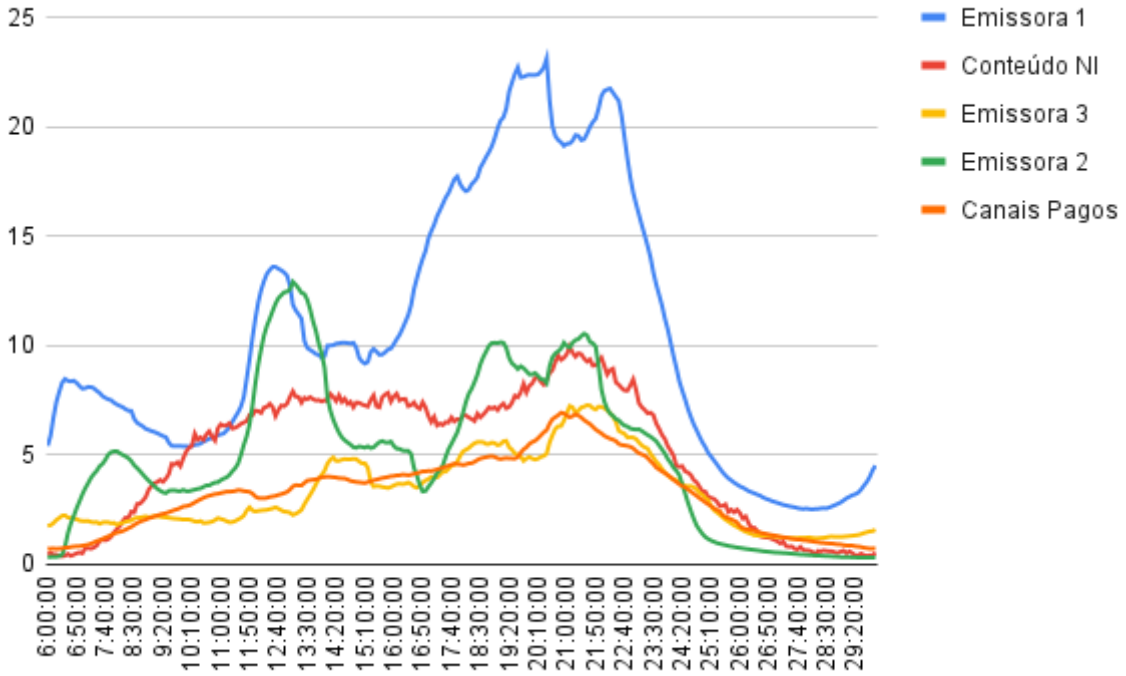
Fizemos um gráfico que reúne todos os domingos do mesmo espaço de tempo (junho de 2020 até junho de 2022), dividimos nos mesmo segmentos do gráfico anterior: todos os dados de Rat% com gêneros, idades e o total de domicílios.

No gráfico de domingo, podemos perceber que a audiência do público de 60+ anos é a menor de todas e se mantém assim durante o dia inteiro, tendo um leve pico no período das 21:45, outro ponto interessante é que o público de 50-59 anos começa o dia com um audiencia bem diferente das outras faixas etárias e por volta de 16:50 começa a se assemelhar com os outros públicos.

Já os públicos de outras faixas etárias se mantêm semelhantes durante o dia inteiro, tendo um aumento de audiência no final da tarde(por volta de 17:40), e começa a ter uma queda de audiência por volta das 22:40. Algo que pode ser percebido nesse intervalo de tempo é que o aumento de audiência no público de 24-35 anos é o mais relevante entre esses públicos.

4.2.6. Comparação entre emissoras

Figura 19 - Gráfico de comparação de audiência entre as emissoras.



Fonte: do próprio autor (2022).

É possível observar uma predominância de audiência da emissora comparada a todas as outras, sendo que a emissora 4 durante o período de almoço tem uma audiência equilibrada com ela, mas durante o resto do dia é menor.

Nesse gráfico também notamos que a emissora 1 e 4 tem um pico de audiência por volta de 12:40, a emissora 1 também tem picos de audiência durante o período da noite mas que se difere, das demais emissoras(por volta de 20:35 e 22:40), a emissora 4 também tem picos durante o período da noite(por volta de 18:55 e 21:25). As demais emissoras têm um constante crescimento de audiência ao longo do dia até as 21:00 e depois disso sua audiência começa a cair e só volta a se recuperar no período da manhã por volta das 10:10.

O aspecto mais marcante desse gráfico é que a emissora 1 não perde quase nunca para nenhuma outra emissora, somente entre 10:10 - 11:50 (contéudo NI) e 13:20 - 14:20 (emissora 2), fora esse ponto, quase não há concorrência para a emissora que foi primeira citada.

4.2.7. Descrição da predição desejada, identificação da sua natureza

A natureza da predição desejada é contínua, uma vez que a saída dos dados será feita através da predição de um score de audiência.

4.3. Preparação dos Dados

Descreva as etapas realizadas para definir os dados e os atributos descritivos dos dados ("features") a serem utilizados. Essa descrição deve ser feita de modo a garantir uma futura reprodução do processo por outras pessoas, e deve conter:

4.3.1. Descrição das manipulações necessárias nos registros e suas respectivas features

Tabelas

Emissora 1,2-Grade Horária (Seg a Sex, Sab, Dom) com Audiência (Seg a Sex, Sab, Dom)

Junção:

Grade Horária com Audiência (Emissora 1,2)

```
#unindo as duas tabelas das quais uma corresponde a informacoes de quem assites e outra com estilo do programa; com base no horario e dia em que passa
new_merged_emissora = pd.merge(rat_seg_sex1, data_seg_sex,how='left',left_on=['Data','Hora Início'], right_on=['Data','Faixa Horária'])
```

Descrição: Juntamos a tabela de Grade Horária com a de audiência para ter acesso à audiência que aquela categoria representa em determinado horário. Assim podendo relacionar a audiência (Rat%) a uma determinada categoria de programa no modelo preditivo.

Junção Seg a Sex com Sab e Dom (Emissora 1,2 e NI)

```
clean_merged3 = pd.concat([clean_seg_sex_streaming,clean_sabado_streaming,clean_domingo_streaming])

clean_merged3['Data'] = clean_merged3['Data'].apply(pd.to_datetime)

clean_merged3.drop_duplicates(inplace=True)

clean_merged3.sort_values(by=['Data','Hora Início'], ascending=True)
```

Descrição: Após a junção da grade horária com a audiência de cada tabela (Seg a Sex, Sab e Dom) e seu tratamento (Agregação e Filtragem) decidimos juntar ela em uma tabela só. Além de juntá-las, organizamos elas em ordem de dia da semana e horário (.sort).

Agregação:

Programação/Categoria (Emissora 1 e 2)

```
#separa uma informacao em duas colunas para melhor leitura da tabela
new_merged_emissora2[['Programa','Categoria']] = new_merged_emissora2['TV '].str.split(' / ', expand=True)
```

Descrição: Para conseguirmos usar separadamente as informações que estavam dentro dessa coluna, dividimos ela em duas. Portanto, usamos o comando ".split" para dividir "Programação/Categoria" pela "/" e criar duas novas colunas.

Data (Emissora 1 e 2)

```
#Separa a coluna data em 3 colunas "Ano, mes e dia"
new_merged_emissora2[['Ano','Mês','Dia']] = new_merged_emissora2['Data'].str.split('-', expand=True)
```

Descrição: Para conseguir utilizar as informações presente na coluna "Data" tivemos que dividi la em três novas colunas "Ano", "Mês" e "Dia"Assim, usamos o "Split" para dividir "Ano-Mês-Dia" pela "-" e criar em três novas colunas para agrupar essas informações.

Filtragem :

Data = Hora início (Emissora 1 e 2)

```
#igualando duas colunas para unir-las
data_sabado['Faixa Horária'] = rat_sabado2['Hora Início']
```

Descrição: Igualamos a coluna "Faixa Horária" com a "Hora Início" para mostrar apenas o horário de início e não o intervalo, assim ficando mais fácil de ordenar por horário outras tabelas.

Exclusão de colunas (Emissora 1,2 e NI)

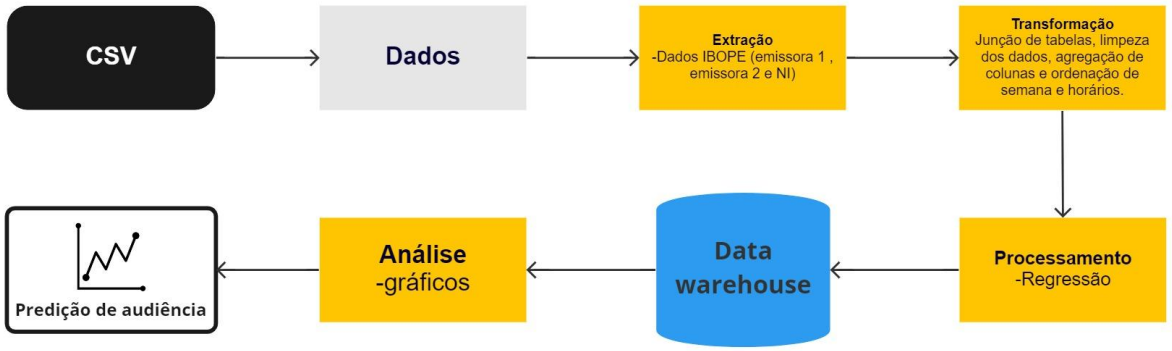
```
#tira as colunas das quais nao vamos usar
clean_domingo_streaming = rat_domingo_streaming.drop(['Total Domicílios | Shr%','AB | Shr%', 'C1 | Shr%', 'C2 | Shr%', 'DE | Shr%', 'Masculino | Shr%',
       'Feminino | Shr%', '4-11 anos | Shr%', '12-17 anos | Shr%',
       '18-24 anos | Shr%', '25-34 anos | Shr%', '35-49 anos | Shr%',
       '50-59 anos | Shr%', '60+ anos | Shr%', 'Total Indivíduos | Rch%',
       'AB | Rch%', 'C1 | Rch%', 'C2 | Rch%', 'DE | Rch%', 'Masculino | Rch%',
       'Feminino | Rch%', '4-11 anos | Rch%', '12-17 anos | Rch%',
       '18-24 anos | Rch%', '25-34 anos | Rch%', '35-49 anos | Rch%',
       '50-59 anos | Rch%', '60+ anos | Rch%', 'Total Indivíduos | Fid%',
       'AB | Fid%', 'C1 | Fid%', 'C2 | Fid%', 'DE | Fid%', 'Masculino | Fid%',
       'Feminino | Fid%', '4-11 anos | Fid%', '12-17 anos | Fid%',
       '18-24 anos | Fid%', '25-34 anos | Fid%', '35-49 anos | Fid%',
       '50-59 anos | Fid%', '60+ anos | Fid%', 'Unnamed: 0',  'Emissora', ], axis=1)
```

Descrição: Como precisamos apenas das colunas com o "Rat%", podemos excluir as outras 40 colunas. Para isso usamos o comando .drop + o nome de cada coluna indesejada. Deixando nossa tabela muito mais leve e precisa.

4.3.2. Como deve ser feita a agregação de registros e/ou derivação de novos atributos

Para serem realizadas as derivações dos atributos na ferramenta, foi utilizada a função Split do Pandas, a qual é responsável por dividir uma string em pequenos pedaços utilizando um "separador", que pode ser escolhido no código.

As manipulações realizadas dos registros que foram disponibilizados pelo IBOPE, foram a divisão das colunas de "programa" e "categoria", que antes eram em conjunto, coluna "PROGRAMA/CATEGORIA", sendo assim ficando separadas, agregando à tabela apenas a sessão categoria. Além disso, em uma outra situação se viu necessária a fragmentação da data em três colunas diferentes, "ano", "mês" e "dia", o que antes era "Data" em apenas uma coluna, utilizando a mesma função.



4.3.3. Identificação das features selecionadas, com descrição dos motivos de seleção.

Inicialmente, foi feita uma seleção das das features de acordo com o consenso do grupo, em relação às quais melhores cabiam para o modelo que seria utilizado. Pensando naquelas que mais agregavam e poderiam ter impacto no output da predição, optou-se por manter as features:

Tabela 1 - Features escolhidas.

Data
Hora de início
Dia da semana
Total Domicílios I Rat%
AB I Rat%
C1 I Rat%
C2 I Rat%
De I Rat%
Masculino I Rat %
4-11 anos I Rat%
12-17 anos I Rat%
18-24 anos I Rat%
25-34 anos I Rat%
35-49 anos I Rat%
50-59 anos I Rat%
60+ anos I Rat%
Programa
Categoria
Feriado
Ano

Dia
Mês

Fonte: do próprio autor (2022).

Sendo o Rat%, a medida média de domicílios que assistiram ao programa (Kantar IBOPE, 2016), acreditou-se para o começo dos testes que seria a porcentagem ideal para ser considerada, deixando assim de lado as medidas de alcance (Shr%) e fidelidade (Fid%).

Assim, considerou-se a relevância do espaço-tempo em que o programa está encaixado, uma vez que de acordo com o horário, dia, mês, ano, dia da semana e se esse é feriado ou não, há uma mudança de pontos atingidos e do perfil de telespectadores, podendo estabelecer uma correlação direta entre essas variáveis e o resultado final.

Além disso, observou-se que os segmentos que cada programa possuía afetava o desempenho no score de audiência das emissoras, juntamente com o perfil do público. Dessa forma foi considerado, cada um dos segmentos referentes às faixas horárias medidas. Ademais, após avaliações, tornou-se claro a relevância de dados sobre o público que estava assistindo, uma vez que esses se mostravam os que mais sofriam alterações de acordo com as demais variáveis.

Neste sentido, após a escolha das features, foi feita uma análise do impacto de cada variável a partir do conceito de multicolinearidade, ou seja, conjunto de características, considerando duas delas que carregam as mesmas ou similares informações, na qual possuem relação linear muito forte. Como consequência, em casos assim, o modelo pode acabar ignorando uma das características, tomando elas como redundantes.

Para conseguir obter os resultados, foram utilizadas as bibliotecas *numpy*, *pandas*, *seaborn* e *matplotlib.pylab* disponíveis para a linguagem Python, a partir do código abaixo:

Figura x - Código para gerar o mapa de multicolinearidade.

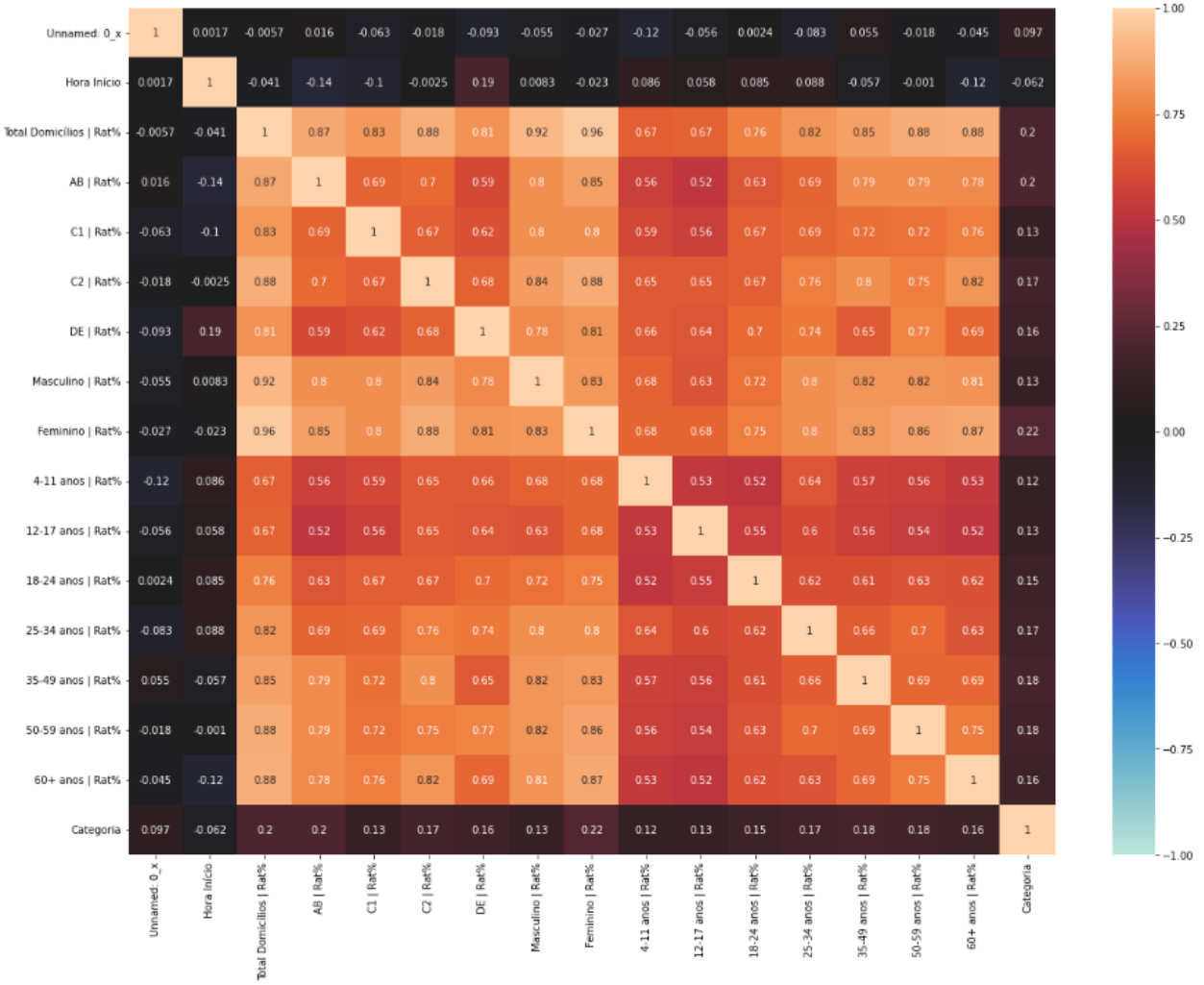
```
## Álgebra Linear
import numpy as np
## Manipulação dos Dados
import pandas as pd
## Visualizações
import seaborn as sns
import matplotlib.pylab as plt
from matplotlib.pylab import rcParams
```

```
[ ]  1 ## Definindo tamanho da figura
     2 rcParams['figure.figsize'] = 20, 15
     3 ## Matriz de correlação
     4 matriz_de_correlacao = clean_merged.corr()
     5 ## Mapa de calor
     6 sns.heatmap(matriz_de_correlacao, annot=True, vmin=-1, vmax=1, center=0)
     7 ## Definindo a posição dos ticks nos eixos
     8 plt.yticks(rotation=360)
     9 plt.xticks(rotation=90)
    10 ## Mostrando a figura
    11 plt.show()
    12
```

Fonte: do próprio autor (2022).

Dessa forma foi observado com a saída da execução, quais das features estariam com dados mais correlacionados, sendo as cores mais claras as associações mais similares, conforme observado no mapa abaixo.

Figura x - Mapa de multicolinearidade.



Fonte: do próprio autor (2022).

4.4. Modelagem

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.

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

Nesta seção você deve incluir as principais referências de seu projeto, para que seu parceiro possa consultar caso ele se interessar em aprofundar.

Utilize a norma ABNT NBR 6023 para regras específicas de referências. Um exemplo de referência de livro:

LUCK, Heloisa. **Liderança em gestão escolar**. 4. ed. Petrópolis: Vozes, 2010.

SOBRENOME, Nome. **Título do livro**: subtítulo do livro. Edição. Cidade de publicação: Nome da editora, Ano de publicação.

MEDIA, Kantar IBOPE. **Dados da audiência nas 15 praças praças regulares com base no ranking consolidado – 04/07 a 10/07.** Disponível em: ‹https://www.kantaribopemedia.com/dados-de-audiencia-nas-15-pracas-regulares-com-base-no-ranking-consolidado-0407-a-1007/›. Acesso em: 21 de agosto de 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.