Formação em Ciência de Dados

Projeto Cidade Ágil

Mortalidade Infantil

Gestão de Indicadores de Saúde

Autores:
Antonio Namur
Eduardo Hitler
Fernanda Castilho
Giane Santana
Silvana Sá

São Paulo, 2020

Sumário

1. Introdução

O Instituto [D'Vinci]³ é uma organização orientada ao ensino que tem como objetivo construir o conhecimento e formar indivíduos para uma sociedade transformadora, onde o estímulo ao pensar, argumentar e responder questões de alta complexidade, têm papéis essenciais. Durante a participação da nossa equipe na formação Ciência de Dados, nos foi atribuído o seguinte desafio: Atender á CIDADE ÁGIL - Consultoria no desenvolvimento de soluções na área de gestão de saúde.

A Cidade Ágil é uma consultoria focada em fornecer um Hub Digital contemplando todas as soluções necessárias ao funcionamento de uma cidade, favorecendo o desenvolvimento integrado e sustentável, de forma a tornar as cidades mais inovadoras, competitivas, atrativas e resilientes, melhorando vidas.

O escopo e objetivo do projeto é evidenciar a situação da saúde nos municípios do estado de São Paulo, desenvolvendo um painel de indicadores que poderá servir de base para a elaboração e aplicação de projetos de políticas públicas mais eficazes.

Os indicadores de desempenho têm como foco, apresentar a figura dos 645 municípios nos espectros da saúde, sociodemografia e infraestrutura. Será possível aos gestores de saúde, realizar projeções que possam dar subsídios e direcionamento aos indicadores que precisam ser melhorados assim como avaliar e monitorar a taxa de mortalidade infantil.

2. Ferramentas Utilizadas

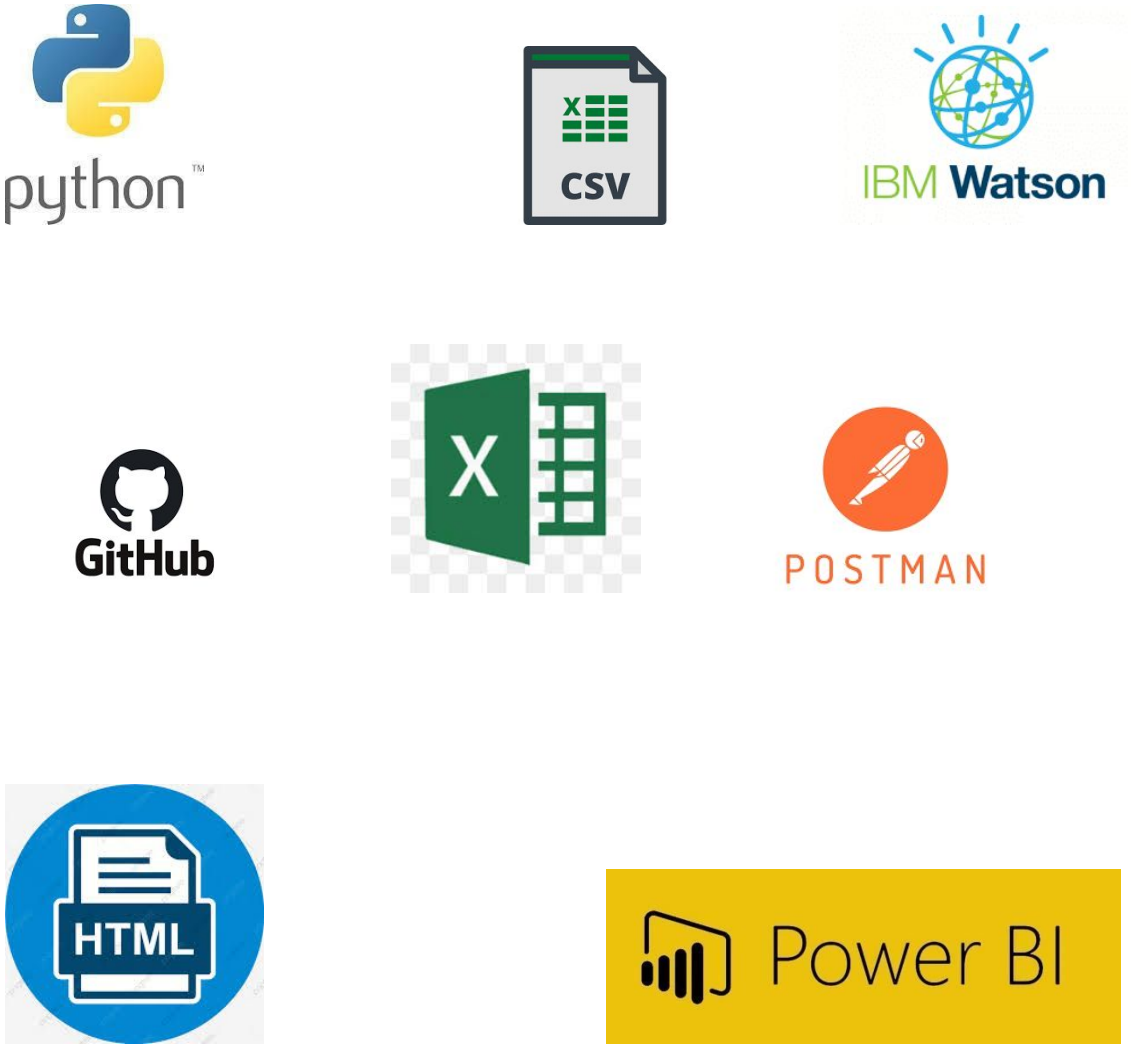


Imagem 1 - Ferramentas utilizadas no projeto de gestão de indicadores

3. Metodologia do Projeto

A seguir, discorreremos sobre a metodologia adotada na execução do projeto.

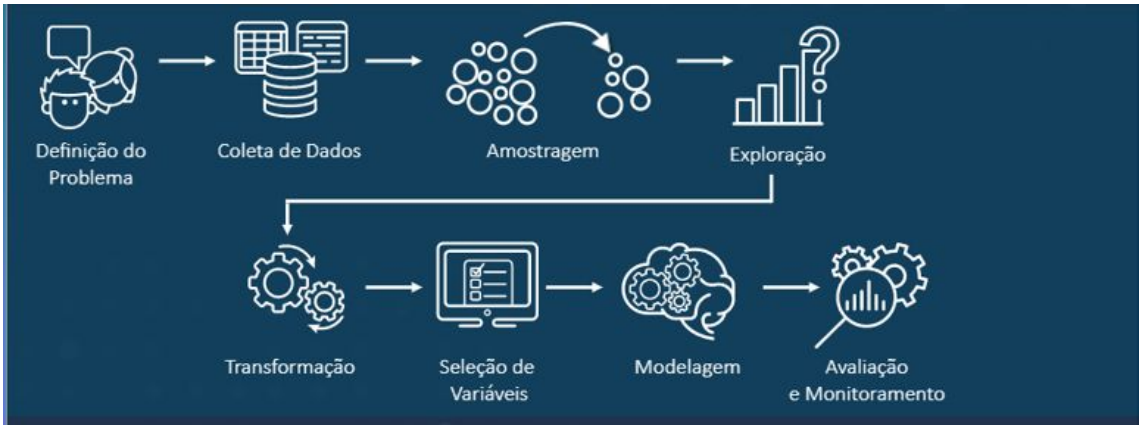


Imagem 2 - Metodologia de ciência de dados - Fonte: SAS

3.1. Definição do problema

A maior dificuldade foi buscar mecanismos para possibilitar que os Gestores e Secretários de Saúde Municipais e Prefeitos adotem mecanismos para medição de indicadores de desempenho, por meio de um conjunto de ferramentas que possibilitem acesso e análise da informação, visando garantir o bem-estar físico, mental e social da população.

3.1.1. Escopo do Projeto

Nossos indicadores têm como foco, apresentar o comportamento das taxas de mortalidade infantil dos municípios do estado de São Paulo. Será possível possibilitar aos gestores de saúde, realizar simulações / projeções que possam dar subsídios e direcionamento dos indicadores que precisam ser melhorados ocasionando uma diminuição gradual da taxa de mortalidade infantil existente.

3.1.2. Desafio

Com a ausência de dados do sistema Hygia, devido à pandemia da Covid-19, optamos por trabalhar apenas com dados públicos.

3.2. Coleta de Dados

Nesta fase, analisamos as fontes das quais podemos extrair os dados, tanto a variável-resposta (target), como as variáveis independentes, ou seja, as variáveis que serão utilizadas para explicar o conjunto de indicadores, dos quais serão construídos os indicadores. Esses dados podem estar em diferentes formatos e vir de diferentes fontes, por isso há a necessidade de consolidação e tratamento (verificação da qualidade dos dados), para que os dados brutos possam ser usados como entrada na modelagem do modelo de regressão que explicaremos mais à frente.

A parte de preparação foi uma das fases mais desafiadoras devido à grande quantidade de informação que vem sendo armazenada. e muito tempo despendido.

Listamos a quantidade de dados que foram coletados, e quais deles foram eleitos para a excussão do projeto. Ao todo elegemos 29 dados (variáveis escolhidas) das quais realizamos a exposição no painel de indicadores.

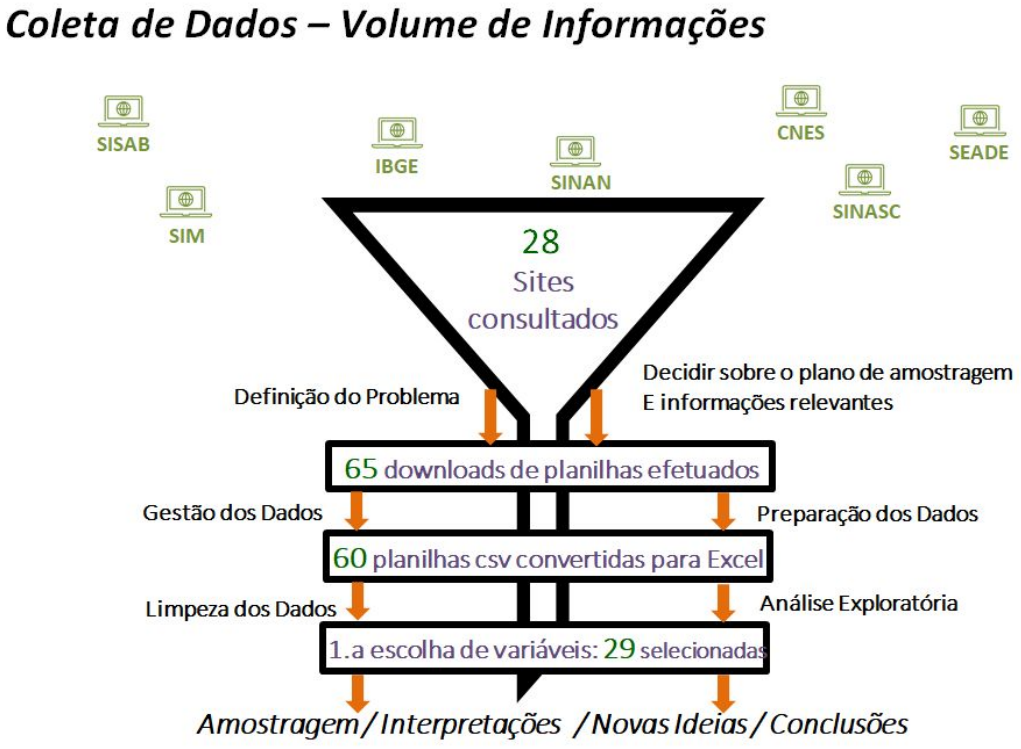


Imagem 3 – Coleta de Dados - Amostragem - Volume de Informações

Os dados coletados são dados públicos e estão disponíveis na internet. Estas informações podem ser consultadas nos sites:

- ✔ CNES - Cadastro Nacional de Estabelecimentos de Saúde
- ✔ IBGE - Instituto Brasileiro Geografia e estatística.
- ✔ SEADE - Fundação Sistema Estadual de Análise de Dados.
- ✔ SIM - Sistema de Informação sobre Mortalidade.
- ✔ SINAN - Sistema de Informação de Agravos e Notificação
- ✔ SINASC - Sistema de Informações de Nascidos Vivos.
- ✔ SISAB -Sistema de Informação em Saúde para a Atenção Básica.
- ✔ SAÚDE BRASIL - Ministério da Saúde.
- ✔ DATASUS - Departamento de Informática do SUS

3.2.1. Dicionário das variáveis

A seguir temos a definição técnica dos principais indicadores utilizados ao longo do projeto. Note que, estão listadas apenas variáveis cujos significados são implícitos e, portanto, precisam de maior explanação.

Mortalidade Infantil

Número de óbitos de menores de um ano de idade, por mil nascidos vivos, na população residente em determinado espaço geográfico, no ano considerado. Estima o risco de morte dos nascidos vivos durante o seu primeiro ano de vida e reflete, de maneira geral, as condições de desenvolvimento socioeconômico e infra-estrutura ambiental, bem como o acesso e a qualidade dos recursos disponíveis para atenção à saúde materna e da população infantil.

Classificação: O valor da taxa é considerado como alto (50 ou mais), médio (20 a 49) e baixo (menos de 20).

Cálculo:

Número de óbito de crianças com menos de 1 ano de idade
÷
(1000 × Nascidos vivos)

Grau Urbanização:

Indica a proporção da população total que reside em áreas urbanas, segundo a divisão político-administrativa estabelecida pelas administrações municipais. Usado para acompanhar o processo de urbanização da população brasileira, em diferentes espaços geográficos e também para subsidiar processos de planejamento, gestão e avaliação de políticas públicas, para adequação e funcionamento da rede de serviços sociais e da infra-estrutura urbana.

Cálculo:

(População urbana residente × 100)
÷
População total residente

Renda per Capita:

É a razão da renda nacional pelo número de pessoas residentes em determinado espaço geográfico, no ano considerado. Considera-se como renda domiciliar per capita a soma dos rendimentos mensais dos moradores do domicílio dividida pelo número de seus moradores. Valores muito baixos assinalam, em geral, a existência de segmentos sociais com precárias condições de vida. É uma característica da unidade domiciliar que é atribuída para cada uma das pessoas nela residente.

Cálculo:

Soma das rendas domiciliares per capita
÷
População total residente

*Onde a **renda domiciliar per capita** é:*

Soma da renda dos moradores
÷
Número de moradores no domicílio

Abastecimento de Água:

Percentual da população residente servida por rede geral de abastecimento, com ou sem canalização domiciliar, em determinado espaço geográfico, no ano considerado. Mede a cobertura de serviços de abastecimento adequado de água à população, por meio de rede geral de distribuição. Expressa as condições socioeconômicas regionais e a priorização de políticas governamentais direcionadas ao desenvolvimento social. Subsidiar análises de risco para a saúde associados a fatores ambientais. Baixas coberturas favorecem a proliferação de doenças transmissíveis decorrentes de contaminação ambiental.

Cálculo:

(População permanente servida por rede geral de abastecimento de água × 100)
÷
População total permanente ajustada para o meio do ano

Esgoto Sanitário:

Percentual da população residente que dispõe de escoadouro de dejetos através de ligação do domicílio à rede coletora ou fossa séptica, em determinado espaço geográfico, no ano considerado. Expressa as condições socioeconômicas regionais e a priorização de políticas governamentais direcionadas ao desenvolvimento social. Analisar variações geográficas e

temporais na cobertura de esgotamento sanitário, identificando situações de desigualdade e tendências que demandem ações e estudos específicos. Subsidiar análises de risco para a saúde associados a fatores ambientais. Baixas coberturas favorecem a proliferação de doenças transmissíveis decorrentes de contaminação ambiental.

Cálculo:

(População permanente servida por rede coletora ou fossa séptica × 100)
÷
População total permanente, ajustada para o meio do ano

Coleta de Lixo:

Percentual da população residente atendida, direta ou indiretamente, por serviço regular de coleta de lixo domiciliar, em determinado espaço geográfico, no ano considerado. Expressa as condições socioeconômicas regionais e a priorização de políticas governamentais direcionadas ao desenvolvimento social. Analisar variações geográficas e temporais na cobertura de serviços de coleta de lixo, identificando situações de desigualdade e tendências que demandem ações e estudos específicos. Subsidiar análises de risco para a saúde associados a fatores ambientais. Baixas coberturas favorecem a proliferação de doenças transmissíveis decorrentes de contaminação ambiental.

Cálculo:

(População permanente servida por serviço regular de coleta de lixo × 100)
÷
População total permanente, ajustada para o meio do ano

IDH - Índice de Desenvolvimento Humano:

Além de computar o PIB per capita, o IDH também leva em conta dois outros componentes: a longevidade e a educação. Para aferir a longevidade, o indicador utiliza números de expectativa de vida ao nascer. O item educação é

avaliado pelo índice de analfabetismo e pela taxa de matrícula em todos os níveis de ensino.

A renda é mensurada pelo PIB per capita. Essas três dimensões têm a mesma importância no índice, que varia de zero a um.

Classificação**:** O IDH está dividido em três grupos distintos: 1) Baixo: de 0 a 0,499; 2) Médio: de 0,500 a 0,799; e 3) Alto: acima de 0,800.

Cálculo**:**

Média das dimensões, Longevidade, Renda e Educação

Índice de Gini:

Valor do Índice de Gini da renda domiciliar per capita das pessoas residentes em determinado espaço geográfico, no ano considerado. Considerou-se como renda domiciliar per capita a soma dos rendimentos mensais dos moradores do domicílio, em reais, dividida pelo número de seus moradores.

Classificação**:** O Coeficiente de Gini consiste em um número entre 0 e 1, onde 0 corresponde à completa igualdade (no caso do rendimento, por exemplo, toda a população recebe o mesmo salário) e 1 corresponde à completa desigualdade (onde uma pessoa recebe todo o rendimento e as demais nada recebem).
O índice de Gini é o coeficiente expresso em pontos percentuais (é igual ao coeficiente multiplicado por 100).

Doses Aplicadas de Vacina

Número de doses de vacinas aplicadas por mil residentes em determinado espaço geográfico, no ano considerado. Considerou-se as doses aplicadas para todos os tipos de imunobiológicos contabilizados pela CGPNI (Coordenação Geral do Programa Nacional de Imunizações - Ministério da Saúde).

Cálculo:

(Número de doses aplicadas $\times$ *1000)*

÷

População total residente

Número de Médicos:

Número de médicos registrados no CNES residentes que atuam em determinado espaço geográfico para cada mil pessoas residentes, no ano considerado. Expressa as condições socioeconômicas regionais e a acessibilidade do atendimento local de saúde. Subsidiar análises de demanda para a saúde associados a fatores socioeconômicos.

Cálculo**:**

(Número de médicos residentes registrados × 1000)

÷

População total residente

Estabelecimentos de Saúde:

Número de estabelecimentos de saúde registrados no CNES existentes para cada mil pessoas residentes em determinado espaço geográfico, no ano considerado. Expressa as condições socioeconômicas regionais e a acessibilidade do atendimento local de saúde. Subsidiar análises de demanda para a saúde associados a fatores socioeconômicos.

Cálculo**:**

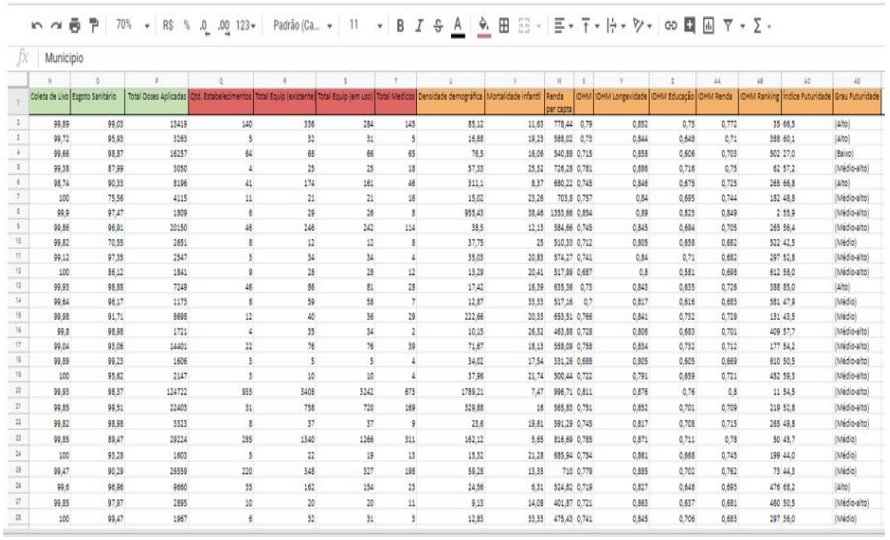
(Número de estabelecimentos de saúde × 1000)

÷

População total residente

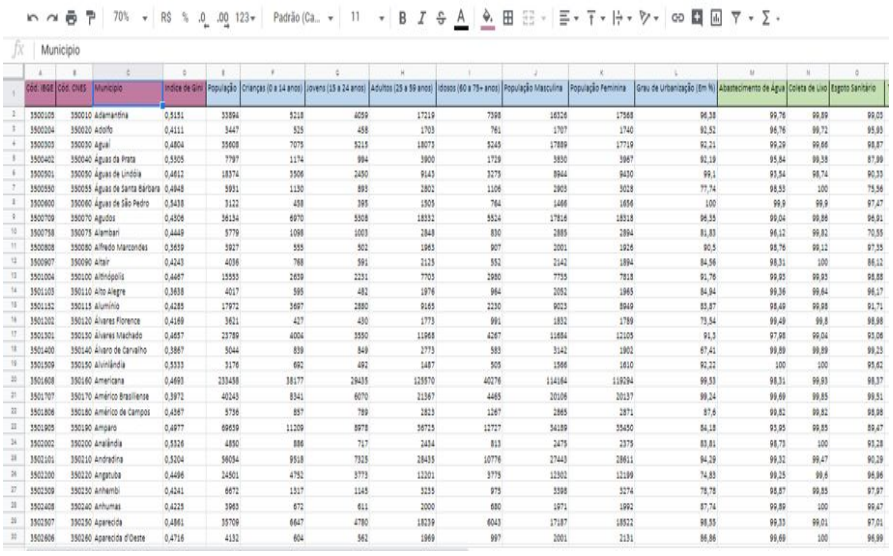
3.3. Amostragem

Nesta fase, após a coleta dos dados, iniciou se a fase de amostragem. Após a seleção dos dados, foi elaborado uma planilha matriz da qual consolidamos os dados para criação do nosso Dataset final. Além disso, nesta fase é definida e dividida a base de estudo em treinamento (para posterior utilização dos modelos), validação (verificar a qualidade dos modelos ajustados e escolher o modelo campeão) e teste (avalia o modelo campeão).



Coleta de Lixo	Esgoto Sanitário	Total Doses Aplicadas	Qtd. Estabelecimentos	Total Equip. existente	Total Equip. em uso	Total Médicos	Densidade demográfica	Mortalidade infantil	Renda per capita	IDHM	IDHM Longevidade	IDHM Educação	IDHM Renda	IDHM Ranking	Índice Futuridade	Grau Futuridade
99,89	99,03	13419	140	336	284	145	85,12	11,63	778,44	0,79	0,852	0,75	0,772	55	66,5	(Alto)
99,72	95,93	3263	5	32	31	5	16,88	18,23	588,02	0,73	0,844	0,648	0,71	388	60,1	(Alto)
99,66	98,87	16257	64	68	66	65	76,5	16,06	540,88	0,715	0,858	0,606	0,703	502	27,0	(Baixo)
99,38	87,99	3050	4	25	25	18	57,33	25,32	726,28	0,781	0,886	0,716	0,75	62	57,2	(Médio-alto)
98,74	90,33	8196	41	174	161	46	311,1	8,37	680,22	0,745	0,846	0,675	0,725	269	66,8	(Alto)
100	75,56	4115	11	21	21	16	15,02	23,26	703,8	0,757	0,84	0,695	0,744	182	48,8	(Médio-alto)
99,9	97,47	1809	6	29	26	8	955,43	38,46	1353,66	0,854	0,89	0,825	0,849	2	55,9	(Médio-alto)
99,86	96,81	20150	46	246	242	114	38,5	12,13	584,66	0,745	0,845	0,694	0,705	265	58,4	(Médio-alto)
99,82	70,55	2651	8	12	12	8	37,75	25	510,33	0,712	0,805	0,658	0,682	522	42,5	(Médio)
99,12	97,35	2547	3	34	34	4	35,05	20,83	574,27	0,741	0,84	0,71	0,682	297	52,8	(Médio-alto)
100	86,12	1841	9	28	28	12	13,39	20,41	517,89	0,687	0,8	0,581	0,688	612	58,0	(Médio-alto)
99,93	98,88	7249	46	86	81	28	17,42	16,39	635,36	0,75	0,843	0,635	0,726	388	55,0	(Alto)
99,64	96,17	1173	8	59	58	7	12,87	33,33	517,16	0,7	0,817	0,616	0,683	581	47,9	(Médio)
99,98	91,71	8698	12	40	36	29	222,66	20,33	653,51	0,766	0,841	0,732	0,729	131	43,5	(Médio)
99,8	98,99	1721	4	35	34	2	10,13	26,32	483,88	0,728	0,806	0,683	0,701	409	57,7	(Médio-alto)
99,04	93,06	14401	22	76	76	39	71,67	18,13	558,09	0,758	0,854	0,732	0,713	177	54,2	(Médio-alto)
99,89	99,23	1606	3	5	5	4	34,02	17,54	551,26	0,688	0,805	0,605	0,669	610	50,5	(Médio-alto)
100	95,62	2147	3	10	10	4	37,96	21,74	500,44	0,722	0,791	0,659	0,721	452	59,3	(Médio-alto)
99,93	98,37	124722	933	3408	3242	673	1789,21	7,47	996,71	0,811	0,876	0,76	0,8	11	54,5	(Médio-alto)
99,85	99,51	22405	31	758	720	169	529,88	16	965,83	0,791	0,852	0,701	0,709	219	52,8	(Médio-alto)
99,82	98,98	3323	8	37	37	9	23,6	18,61	591,28	0,745	0,817	0,708	0,715	265	48,8	(Médio-alto)
99,85	99,47	39224	285	1340	1266	311	162,12	9,69	816,69	0,785	0,871	0,711	0,78	90	45,7	(Médio)
100	93,28	1603	5	22	18	13	13,32	21,28	665,94	0,754	0,861	0,668	0,745	199	44,0	(Médio)
99,47	90,29	26559	220	348	327	198	59,28	13,35	710	0,779	0,885	0,702	0,762	73	44,3	(Médio)
99,6	96,96	9660	35	162	154	25	24,56	6,31	924,83	0,719	0,827	0,648	0,693	476	68,2	(Alto)
99,85	97,97	2895	10	20	20	11	9,13	14,08	401,87	0,721	0,863	0,637	0,681	460	50,5	(Médio-alto)
100	99,47	1967	6	32	31	5	12,83	33,33	475,43	0,741	0,845	0,706	0,683	297	36,0	(Médio)

Imagem 4 – Data Set Amostragem.



Cód. IBGE	Cód. CNES	Município	Índice de Gini	População	Crianças (0 a 14 anos)	Jovens (15 a 24 anos)	Adultos (25 a 59 anos)	Idosos (60 a 75+ anos)	População Masculina	População Feminina	Grau de Urbanização (Em %)	Abastecimento de Água	Coleta de Lixo	Esgoto Sanitário
3500105	350010	Adamantina	0,5151	35894	5218	4059	17219	7398	16326	17568	96,38	99,76	99,89	99,05
3500204	350020	Adolfo	0,4111	3447	525	458	1703	761	1707	1740	92,52	96,76	99,72	95,93
3500303	350030	Aguaí	0,4804	35608	7079	5215	18073	5240	17889	17719	92,21	99,29	99,66	98,87
3500402	350040	Águas da Prata	0,5305	7797	1174	984	3900	1729	3850	3967	92,19	95,84	99,38	87,99
3500501	350050	Águas de Lindóia	0,4612	18374	3506	2450	9143	3275	8944	9430	99,1	93,54	98,74	90,33
3500550	350055	Águas de Santa Bárbara	0,4948	5931	1130	893	2802	1106	2903	3028	77,74	98,53	100	75,56
3500600	350060	Águas de São Pedro	0,5438	3122	458	395	1905	764	1466	1656	100	99,9	99,9	97,47
3500709	350070	Agudos	0,4306	36134	6970	5308	18332	5524	17816	18318	96,35	99,04	99,86	96,91
3500758	350075	Alambari	0,4448	5779	1098	1003	2848	830	2889	2894	81,83	96,12	99,82	70,55
3500808	350080	Alfredo Marcondes	0,3659	3927	555	502	1965	907	2001	1926	90,5	98,76	99,12	97,35
3500907	350090	Altair	0,4243	4036	768	591	2125	552	2142	1894	84,56	98,31	100	86,12
3501004	350100	Altinópolis	0,4467	15555	2659	2231	7705	2960	7755	7818	91,76	99,95	99,95	98,88
3501103	350110	Alto Alegre	0,3638	4017	595	482	1976	964	2052	1965	84,94	99,36	99,64	96,17
3501152	350115	Alumínio	0,4285	17972	3497	2880	9165	2230	9023	8949	83,87	98,49	99,98	91,71
3501202	350120	Álvares Florence	0,4169	3621	427	430	1773	991	1832	1789	73,54	99,49	99,8	98,98
3501301	350130	Álvares Machado	0,4457	23789	4004	3550	11968	4267	11684	12105	91,3	97,98	99,04	93,06
3501400	350140	Álvaro de Carvalho	0,3867	5044	839	849	2773	583	3142	1902	67,41	99,89	99,89	99,23
3501509	350150	Alvinlândia	0,5333	3176	692	492	1487	505	1566	1610	92,22	100	100	95,62
3501608	350160	Americana	0,4693	233458	38177	29435	125870	40276	114164	119294	99,53	98,31	99,93	98,37
3501707	350170	Américo Brasiliense	0,3972	40243	8341	6070	21367	4465	20106	20137	99,24	99,69	99,85	99,51
3501806	350180	Américo de Campos	0,4367	5736	857	789	2823	1267	2865	2871	87,6	99,82	99,82	98,98
3501905	350190	Amparo	0,4977	69659	11209	8978	36725	12727	34189	35450	94,18	93,95	99,85	89,47
3502002	350200	Analândia	0,5326	4850	886	717	2434	813	2479	2379	83,81	98,73	100	93,28
3502101	350210	Andradina	0,5204	56054	9518	7925	28435	10776	27443	28611	94,29	99,32	99,47	90,29
3502200	350220	Angatuba	0,4496	24501	4752	3773	12201	3775	12302	12199	74,83	99,25	99,6	96,96
3502309	350230	Anhembi	0,4241	6672	1317	1145	3255	975	3398	3274	78,78	98,87	99,85	97,97
3502408	350240	Anhumas	0,4225	3963	672	611	2000	680	1971	1992	87,74	99,89	100	99,47
3502507	350250	Aparecida	0,4861	35709	6647	4780	18239	6043	17187	18522	98,55	99,33	99,01	97,01
3502606	350260	Aparecida d'Oeste	0,4716	4132	604	562	1969	997	2001	2131	86,86	99,69	100	96,99

Imagem 5 – Data Set Amostragem.

3.4. Exploração

Aqui é onde abrangemos tanto a parte da estatística descritiva como a diagnóstica, ou seja, exploramos os dados procurando tendências ou anomalias inesperadas para obter uma melhor compreensão e assim entender o que, e por que, aconteceu. Precisamos saber com o que estamos trabalhando. Verificar as distribuições das variáveis que pretendemos usar e as relações bivariadas entre todas as variáveis foi por onde começamos.

- ✔ Mapa de calor – Correlações entre as variáveis.
- ✔ Seleção de 10 variáveis – Demográficas, Socioeconômicas e de Saúde.
- ✔ Análise de distribuições.

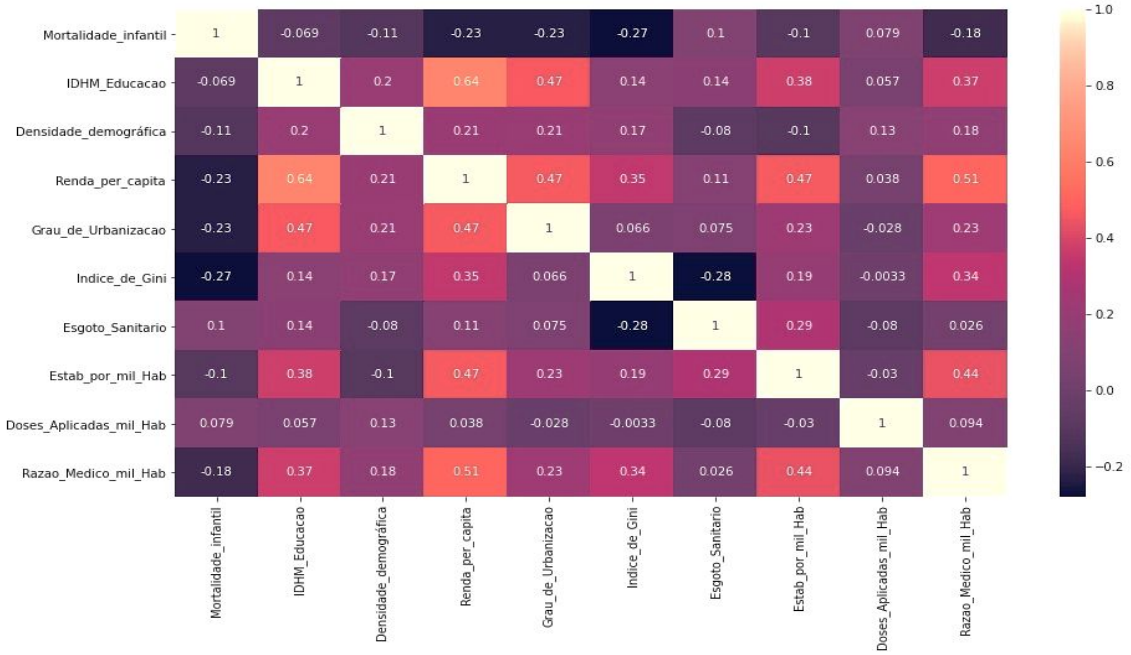


Imagem 6 – Exploração – Mapa de Calor.

Obs.:Identificar a correlação entre as variáveis foi premissa inicial para que pudéssemos seguir para a próxima fase de modelagem.

3.5. Transformação

Ao examinar os dados, na fase de exploração, encontramos a necessidade de criar, transformar, excluir ou combinar variáveis a fim de construir o modelo a ser executado na fase de modelagem.

Nesta etapa seguimos um passo a passo que nos facilitou no direcionamento para conclusão dessa etapa.

- ✔ Identificação e Tratamento dos Outliers
- ✔ Tratamento de missing (valores faltantes)
- ✔ Exclusão de variáveis redundantes (variáveis independentes altamente correlacionadas) e irrelevantes (variável independente com baixa correlação com a resposta, ou seja, não contribui para sua previsão / ou modelo).

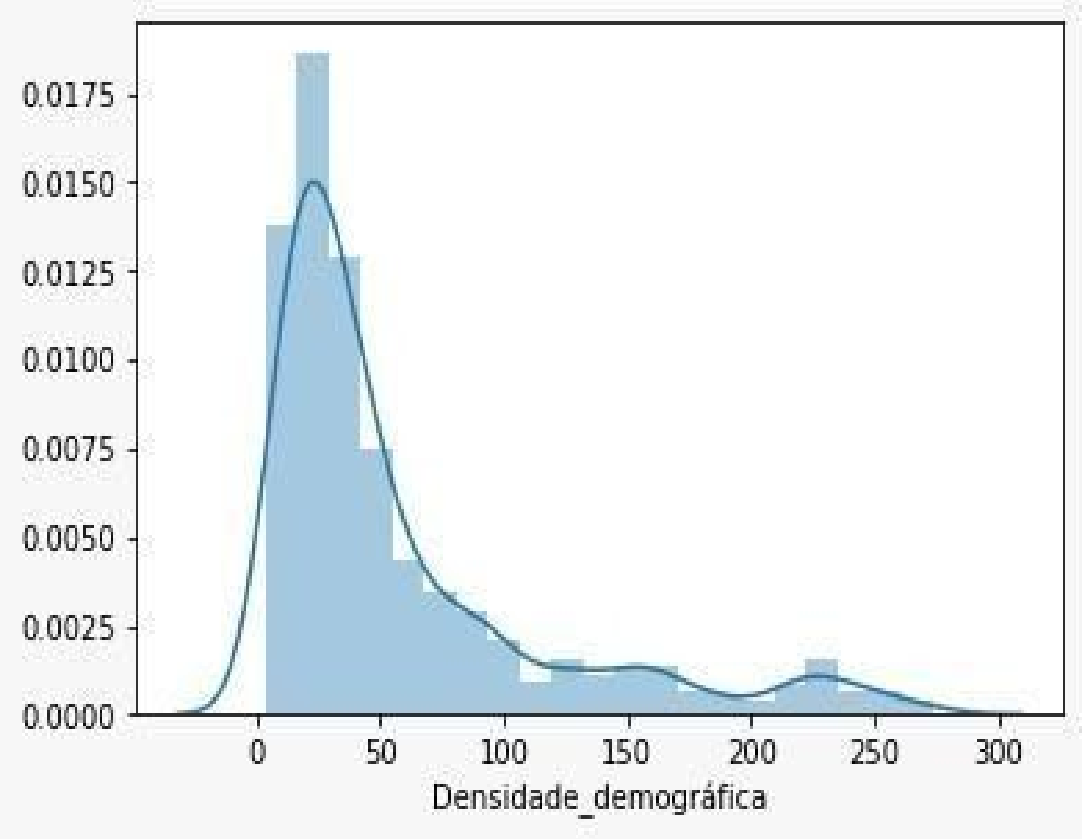


Imagem 7 – Tratamento dos Outliers – IBM Watson Studio

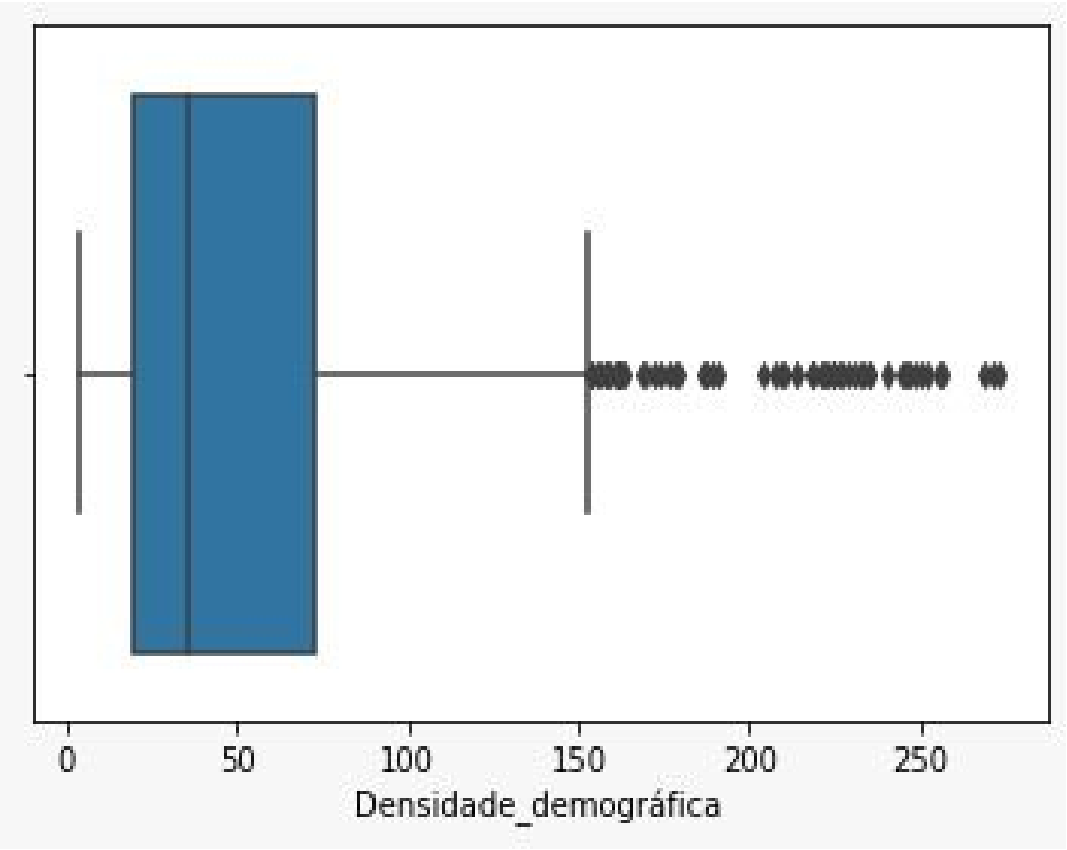


Imagem 8 – Tratamento dos Outliers – IBM Watson Studio

3.6. **Seleção de Variáveis**

Durante a fase de exploração e transformação dos dados, manteve-se a escolha das variáveis socioeconômicas e de saúde, realizando a seleção das variáveis com maior precisão. Desta forma reduzimos o número de variáveis para 4.

3.7. Modelagem

Durante a etapa de modelagem foi executado duas análises: uma análise Não Supervisionada a Clusterização, e posterior a Supervisionada, da qual originou o nosso modelo de regressão.

3.7.1. Análise Não Supervisionada

Realizamos a análise não supervisionada utilizando o método de Clusterização, com o objetivo de agrupar os municípios com que apresentassem semelhanças dentre suas variáveis socioeconômicas e de saúde.

Por que optamos por esse modelo? Optamos pelo método não supervisionado na 1º fase do projeto para que pudéssemos segmentar os 645 municípios do estado de São Paulo, com o objetivo de agrupar os mesmos de uma forma mais homogênea. Com os clusters foi possível explorar os dados e conhecer as similaridades dos municípios do estado de São Paulo com o objetivo de extrair informações das quais poderiam subsidiar uma análise supervisionada na segunda fase do projeto.

Imagem 9 - IBM WATSON STUDIO – MODELER CLUSTER

3.7.1.1. A escolha das Variáveis

Dentre as variáveis selecionadas durante a fase de exploração dos dados, escolhemos as seguintes para usar no modelo:

- ✔ Mortalidade Infantil
- ✔ IDH de Educação
- ✔ Estabelecimentos por mil habitantes
- ✔ Médicos por mil habitantes

3.7.1.2. Algoritmo de Clusterização: K-Means

O algoritmo gera “k” cluster (grupos) encontrando o centróide mais próximo (utilizamos a métrica da distância Euclidiano) dentre as variáveis e atribui o ponto encontrado a um determinado cluster. Os centróides são atualizados sempre tomando o valor médio de todos os pontos naquele cluster. Para este método usamos por convenção apenas valores numéricos para o cálculo da distância. O procedimento foi bem sucedido pois, os dados foram separados organicamente podendo assim ser rotulados e posteriormente usados como referência para análises posteriores.

- ✔ Resultado da Clusterização: 3 grupos.
- ✔ Variável com maior importância dentre as utilizadas: Mortalidade Infantil.

3.7.1.3. Análise do Resultado

O resultado da clusterização nos permitiu fazer algumas observações e inferências a respeito de nossos dados. Pudemos observar que os grupos representavam de maneira clara,três classes específicas de municípios, sendo elas:

- ✔ Grupo 1 – Em estado Satisfatório
- ✔ Grupo 2 – Em estado de Alerta
- ✔ Grupo 3 – Em estado de Atenção

O grupo 1 contém os municípios que apresentam índices mais satisfatórios em comparação aos outros, possui o maior índice de educação e de desenvolvimento social, o menor índice de taxa de mortalidade infantil e apresenta uma boa taxa de nº de médicos por mil habitantes. O grupo 2 apresenta índices medíocres em todos os critérios de saúde, e tem notavelmente os índices mais baixos de educação, ele também é o maior dos 3 grupos, representando quase metade dos municípios de São Paulo. Em contraste, o grupo 3 é o menor dos grupos e também o mais preocupante, poderíamos classificá-lo como municípios em risco por terem os piores índices de saúde e uma altíssima mortalidade infantil.

3.7.2. Análise Supervisionada

Realizamos a análise supervisionada utilizando o método de Regressão, com o objetivo de inferir de que maneira o índice de mortalidade infantil dos municípios se comporta em função dos demais indicadores selecionados.

Por que optamos por esse modelo? Optamos pelo método supervisionado como 2º fase do projeto para que pudéssemos tirar o máximo proveito dos resultados obtidos com a análise feita previamente. O objetivo agora é criar um modelo que possa simular o comportamento da variável alvo dentro de cada um dos grupos identificados gerando valor e novos *insights* sobre nosso dados.

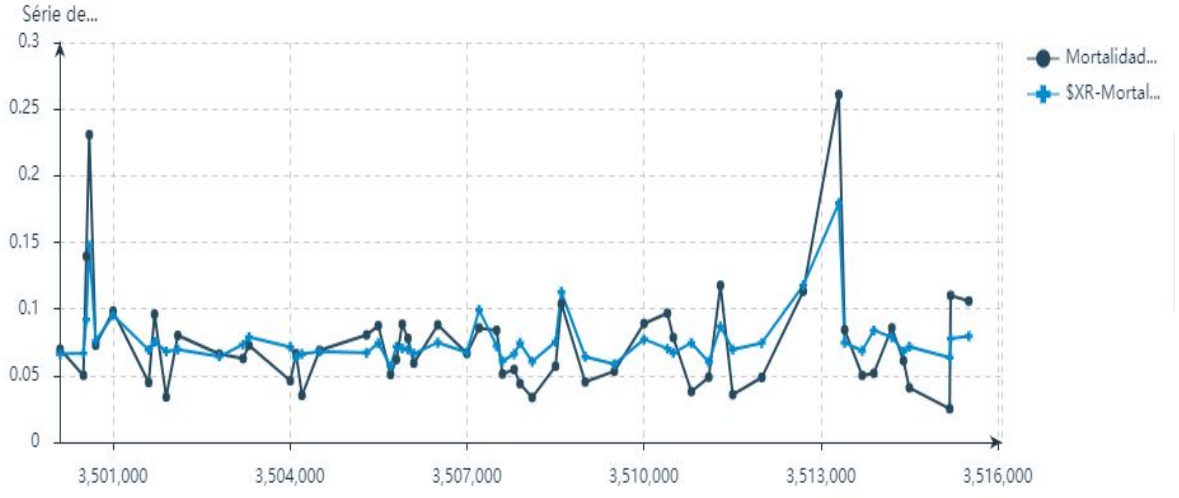


Imagem 10 - IBM WATSON STUDIO – Importance Feature

3.7.2.1.A escolha das Variáveis

Dentre as variáveis selecionadas durante a fase de exploração dos dados, escolhemos as seguintes para usar no modelo:

Variável dependente (target)

- ✔ Mortalidade Infantil

Variáveis independentes (inputs)

- ✔ IDH de Educação
- ✔ Doses de vacina aplicadas por mil habitantes
- ✔ Cobertura de abastecimento de água
- ✔ Cobertura de coleta de lixo
- ✔ Cobertura de esgoto sanitário
- ✔ Grau de Urbanização
- ✔ Renda per capita
- ✔ Estabelecimentos por mil habitantes
- ✔ Médicos por mil habitantes

3.7.2.2. Algoritmo utilizado: XG Boost Tree

O XGBoost é usado para problemas de aprendizado supervisionado, nos quais usamos os dados de treinamento com um alto número de variáveis independentes. Para isso o algoritmo usa o método de árvores de decisão que são estruturas de dados formadas por um conjunto de elementos que armazenam informações, com o diferencial de possuir embutido um sistema próprio poderoso de otimização.

Fizemos o particionamento dos dados em 90% para treino e 10% para teste, e alcançamos bons resultados.

Resultado da regressão para cada grupo:

- ✔ **Grupo 1** (Em estado **Satisfatório**) – **95%** de acurácia
- ✔ **Grupo 2** (Em estado de **Alerta**) – **98%** de acurácia
- ✔ **Grupo 3** (Em estado de **Atenção**) – **86%** de acurácia

3.8. Avaliação e Produção

Após a produção de um modelo funcional, é necessário o desenvolvimento de uma interface onde o usuário final do produto possa interagir ativamente com os dados assim como extrair destes valiosos *insights.*

3.8.1. Deploy do modelo

Nosso primeiro passo para garantir essa entrega foi realizar o deploy do modelo na nuvem, utilizando o próprio servidor da IBM Watson Studio. Dessa maneira nosso modelo pode ser chamado como API em qualquer plataforma com interface Python, JavaScript ou Curl.

Teste de API do Modelo

In [ ]:
```
import requests

# Paste your Watson Machine Learning service apikey here
# Use the rest of the code sample as written
apikey = "c-MaC8NEy0LDY_wHNKyx0LhwnEei7bnAYFnXmQgxK1jp"

# Get an IAM token from IBM Cloud
url     = "https://iam.bluemix.net/oidc/token"
headers = { "Content-Type" : "application/x-www-form-urlencoded" }
data    = "apikey=" + apikey + "&grant_type=urn:ibm:params:oauth:grant-type:apikey"
IBM_cloud_IAM_uid = "bx"
IBM_cloud_IAM_pwd = "bx"
response  = requests.post( url, headers=headers, data=data, auth=( IBM_cloud_IAM_uid, IBM_cloud_IAM_pwd ) )
iam_token = response.json()["access_token"]

ml_instance_id = "ba659903-440e-4610-850f-115a8070c983"
iam_token
```

Imagem 11 - IBM WATSON STUDIO – Deploy do Modelo

3.8.2. Monitoramento e Visualização dos dados

A etapa final do projeto consiste em um trabalho puro de Data Visualization. Garantimos que os dados estivessem expostos da forma mais inteligível possível, acompanhado por suas respectivas métricas de avaliação, usabilidade simples e intuitiva. Dentro do dashboard demos maior ênfase aos Clusters que obtivemos no modelo de análise não supervisionado e as telas se propõem a contar uma história com protagonista bem definido e com começo, meio e fim .

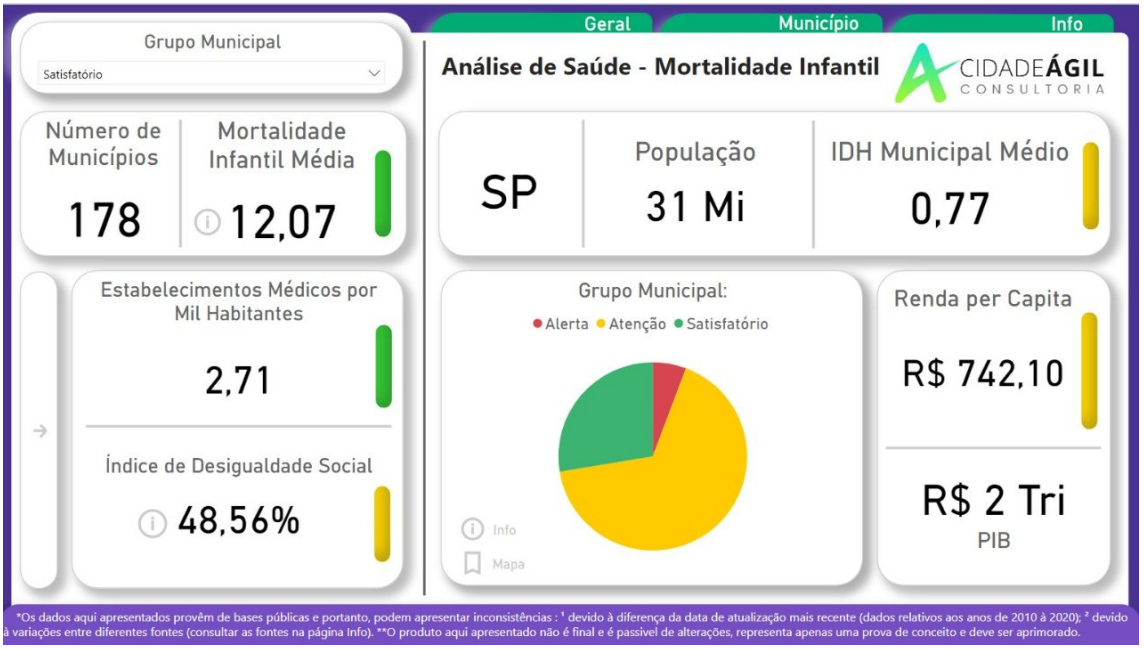


Imagem 12 - Painel para visualização dos Dados de Saúde em Power BI - Tela Geral

Acesse o painel em: http://idsimulador.herokuapp.com/

3.8.3. Simulador de indicadores

Junto ao painel de indicadores, entregamos também um simulador do Índice de Mortalidade Infantil. Trata-se de uma interface interativa onde o usuário pode manipular diversos indicadores como Vacinação, Número de Médicos ou Renda per Capita, e simultaneamente observar como isso interfere na taxa do município. O simulador foi feito com base nos modelos de regressão e trata com a devida diferença cada um dos grupos delimitados.

O produto final ainda precisa ser polido pois ao lidar com municípios cujo os valores estão muito distantes da média do seu grupo, o modelo se confunde e vemos flutuações em seus resultados. Também acreditamos que o modelo esteja em certa parcela com *overfitting* o que pode tornar uma análise tendenciosa, mas acreditamos que com maior riqueza de dados e tempo disposto ao projeto o valor desta entrega se sustenta.

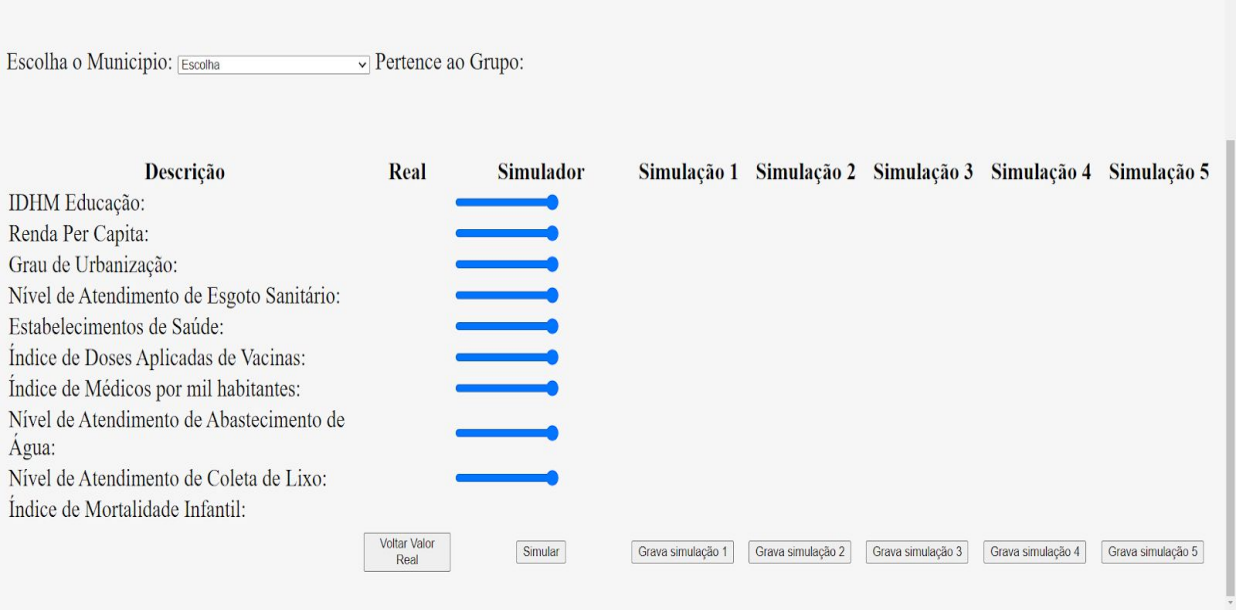


Imagem 13 - Tela de Simulação de Cálculo de Taxa de Mortalidade de Municípios de São Paulo em HTML

4. Referências Bibliográficas

4.1. Principais Fontes de Informação da saúde do Brasil

Quanto às fontes de informação, o Saúde Brasil tenta explorar, em sua máxima potencialidade, os dados originados de sistemas nacionais de informações em saúde, distribuídos por diversos sites, como:

- ★ Sistema de Informações sobre Mortalidade (SIM)
- ★ Sistema de Informações de Nascidos Vivos (Sinasc)
- ★ Sistema de Informação de Agravos e Notificação (Sinan)
- ★ Sistema de Informações Hospitalares (SIH)
- ★ Sistema de Informações Ambulatoriais do SUS (SIA/ SUS)
- ★ Departamento de Informática do SUS (DATASUS)
- ★ Sistema de Informação de Vigilância da Qualidade da Água para Consumo Humano (Sisagua)
- ★ Sistema Nacional de Gestão da Assistência Farmacêutica (Hórus)
- ★ Relação Nacional de Medicamentos Essenciais (Rename)
- ★ Rede Própria do Programa Farmácia Popular do Brasil (PFPB)
- ★ Câmara de Regulação do Mercado de Medicamentos (Cmed)
- ★ Banco de Preços em Saúde (BPS)
- ★ Sistema Integrado de Administração de Serviços Gerais (Siasg)

e demais bases de dados produzidas nos serviços e para os serviços de saúde, assim como inquéritos e utilizam-se, também, informações de outras fontes, tais como:

- ★ Cadastro Nacional de Estabelecimentos de Saúde (Cnes)
- ★ Cadastro de Veículos do Departamento Nacional de Trânsito (Denatran/Ministério das Cidades)
- ★ informações demográficas provenientes dos censos populacionais, Pesquisas Nacionais por Amostra de Domicílios (Pnad)
- ★ Pesquisa Nacional de Saúde (PNS) do Instituto Brasileiro de Geografia e Estatística (IBGE) e ainda
- ★ Instituto de Pesquisa Econômica e Aplicada (Ipea).

4.2. Sites Consultados

Abaixo, listamos as principais páginas públicas de saúde que utilizamos para o nosso estudo e execução do painel e simulador:

1. http://cnes.datasus.gov.br/
2. https://www.ibge.gov.br/
3. https://datasus.saude.gov.br/
4. https://www.seade.gov.br/
5. https://sim.saude.gov.br/
6. https://portalsinan.saude.gov.br/
7. https://sinasc.saude.gov.br/
8. https://sisab.saude.gov.br/
9. https://nacoesunidas.org/pos2015/ods3/

5. Estudos Paralelos

Estudo das variáveis que podem influenciar no índice da taxa de mortalidade infantil.

“De modo geral, expressa o desenvolvimento socioeconômico e a infraestrutura ambiental precários, que condicionam a desnutrição infantil e as infecções a ela associadas. O acesso e a qualidade dos recursos disponíveis para atenção à saúde materno-infantil são também determinantes da mortalidade nesse grupo etário.1 A mortalidade em menores de 5 anos (ou mortalidade na infância) constitui um indicador-chave na avaliação da situação de saúde da população.”

(Texto extraído do saude_brasil_2017_análise_situacao_saude_desafios_objetivos_desenvolvimento_sustetantavel.)

Há consistente tendência de redução da mortalidade infantil em todas as regiões brasileiras, o que reflete a melhoria nas condições de vida, o declínio da fecundidade e o efeito de intervenções públicas nas áreas de saúde, saneamento e educação da mãe, entre outros aspectos.

Metas da ONU, voltada à saúde e bem estar, a serem cumpridas até 2030. Destacamos a meta 3, conforme descrevemos abaixo:

Meta 3. Assegurar uma vida saudável e promover o bem-estar para todas e todos, em todas as idades.

3.1 Até 2030, reduzir a taxa de mortalidade materna global para menos de 70 mortes por 100.000 nascidos vivos.

3.2 Até 2030, acabar com as mortes evitáveis de recém-nascidos e crianças menores de 5 anos, com todos os países objetivando reduzir a mortalidade neonatal para pelo menos 12 por 1.000 nascidos vivos e a mortalidade de crianças menores de 5 anos para pelo menos 25 por 1.000 nascidos vivos

3.3 Até 2030, acabar com as epidemias de AIDS, tuberculose, malária e doenças tropicais negligenciadas, e combater a hepatite, doenças transmitidas pela água, e outras doenças transmissíveis.

3.4 Até 2030, reduzir em um terço a mortalidade prematura por doenças não transmissíveis via prevenção e tratamento, e promover a saúde mental e o bem-estar.

3.5 Reforçar a prevenção e o tratamento do abuso de substâncias, incluindo o abuso de drogas entorpecentes e uso nocivo do álcool.

3.6 Até 2020, reduzir pela metade as mortes e os ferimentos globais por acidentes em estradas.

3.7 Até 2030, assegurar o acesso universal aos serviços de saúde sexual e reprodutiva, incluindo o planejamento familiar, informação e educação, bem como a integração da saúde reprodutiva em estratégias e programas nacionais.

3.8 Atingir a cobertura universal de saúde, incluindo a proteção do risco financeiro, o acesso a serviços de saúde essenciais de qualidade e o acesso a medicamentos e vacinas essenciais seguros, eficazes, de qualidade e a preços acessíveis para todos.

3.9 Até 2030, reduzir substancialmente o número de mortes e doenças por produtos químicos perigosos, contaminação e poluição do ar e água do solo.

3.a Fortalecer a implementação da Convenção-Quadro para o Controle do Tabaco em todos os países, conforme apropriado.

3.b Apoiar a pesquisa e o desenvolvimento de vacinas e medicamentos para as doenças transmissíveis e não transmissíveis, que afetam principalmente os países em desenvolvimento, proporcionar o acesso a medicamentos e vacinas essenciais a preços acessíveis, de acordo com a Declaração de Doha, que

afirma o direito dos países em desenvolvimento de utilizarem plenamente as disposições do acordo TRIPS sobre flexibilidades para proteger a saúde pública e, em particular, proporcionar o acesso a medicamentos para todos.

3.c Aumentar substancialmente o financiamento da saúde e o recrutamento, desenvolvimento e formação, e retenção do pessoal de saúde nos países em desenvolvimento, especialmente nos países menos desenvolvidos e nos pequenos Estados insulares em desenvolvimento.

3.d Reforçar a capacidade de todos os países, particularmente os países em desenvolvimento, para o alerta precoce, redução de riscos e gerenciamento de riscos nacionais e globais de saúde.

6. Conclusão

Há uma constatação de que a mortalidade infantil na maioria dos municípios do estado de São Paulo é de caráter ordinário e essa afirmação pode ser estendida a maioria dos índices socioeconômicos e de saúde. Em contraste, foi identificado um seleto grupo de municípios que se caracterizam por ter taxas elevadas de mortalidade infantil, e que apresentam particularidades similares entre si, em especial são municípios com baixa renda e baixa população. Também foi observado que o índice de desigualdade social é mais elevado em municípios com alta densidade demográfica, como por exemplo no município de São Paulo.

Os Índices socioeconômicos de modo geral expressam alta correlação com a infraestrutura dos municípios e se refletem nas condições de saúde pública.

Nosso projeto consiste em um diagnóstico, porém não pudemos evitar o anseio por criar algo tangível e que pudesse ser mudado de imediato por aqueles que tomam decisões de alto impacto. Com isso em mente desenvolvemos um modelo que simula o comportamento da mortalidade infantil baseado em uma gama variada de índices, tudo isso em tempo real. O produto está longe de finalizado, porém seus benefícios se mostram promissores.

"*Inteligência é a capacidade de absorver informação em tempo real, fazer perguntas que façam sentido, ter boa memória, traçar pontes entre assuntos que não parecem estar relacionados e inovar ao fazer essas conexões.*" – Bill Gates, Microsoft.