MODELOS DE APRENDIZADO

Floresta Aleatória, mapa de calor, correlação linear, análise importância de variáveis, presença de zeros e particularidade das bases de dados.

1. BANCO DE DADOS UTILIZADO PARA ANÁLISE

Explorando Áreas de Ponderação em Minas Gerais.

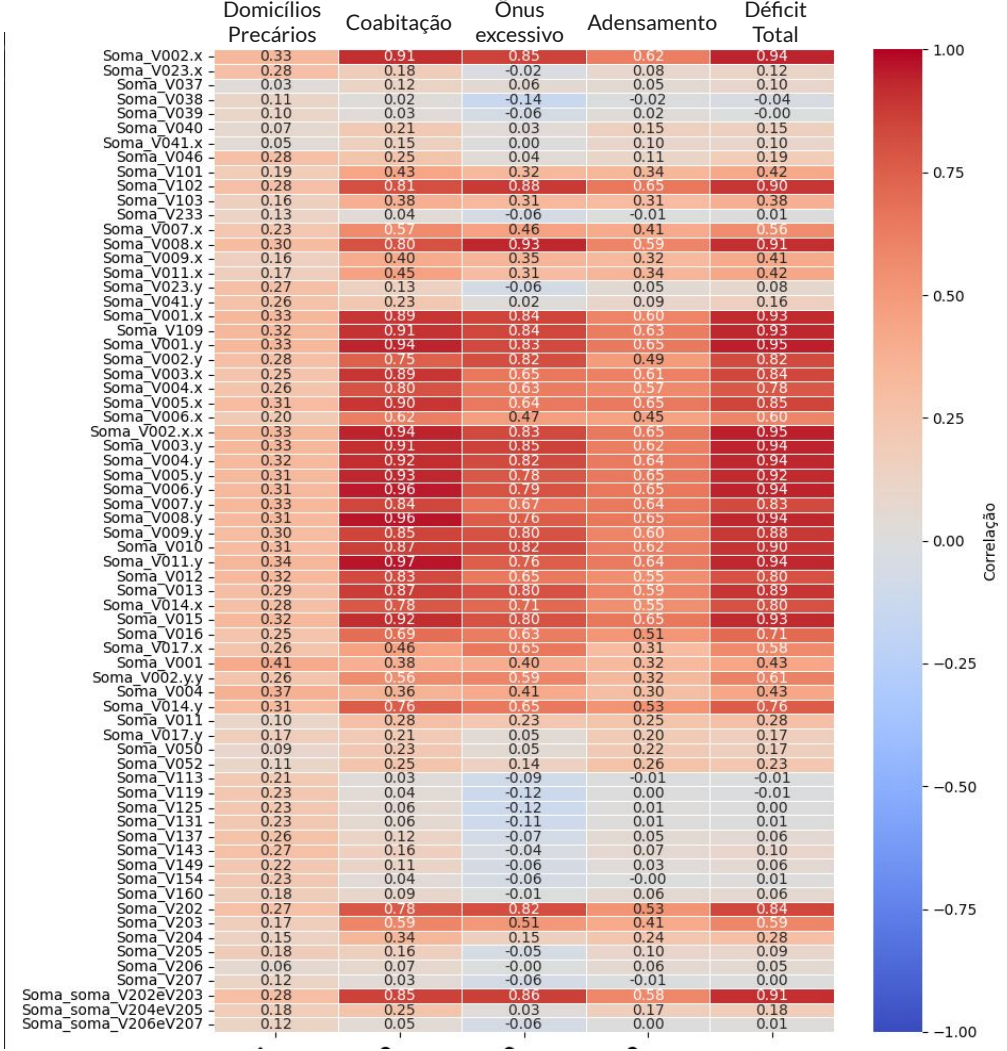
Soma_V002.x	Soma_V023.x	Soma_V037	Soma_V038	Soma_V039	Soma_V040	Soma_V041.x	Soma_V046	Soma_V101	Soma_V102	Soma_V103	Soma_V233	Soma_V007.x	Soma_V008.x	Soma_V009.x	Soma_V011.x	Soma_V023.y	Soma_V041.y	So
1473	0	4	20	0	8	0	0	32	344	10	0	105	968	32	14	0	0	48
6659	21	9	73	2	16	0	10	307	1578	66	18	1045	4942	175	26	44	16	26
2198	3	34	66	1	6	7	8	28	472	13	3	96	1690	52	9	6	13	66
832	2	817	8	1	1	1	4	24	154	2	0	71	486	6	1	5	7	32
1472	4	113	670	10	38	8	3	11	205	3	21	35	612	10	41	7	6	59
2006	17	443	236	4	64	3	19	21	384	11	10	90	1216	67	9	50	50	89
536	0	17	1	0	1	0	1	3	104	5	0	10	288	28	0	0	4	14
791	0	0	4	0	0	0	0	42	123	2	0	145	370	8	0	0	0	22
4133	32	15	94	1	25	2	76	147	706	13	6	539	2357	42	85	96	241	15
2452	28	6	61	0	10	1	29	116	159	14	11	372	526	53	16	94	92	10
6567	12	82	274	5	22	3	26	336	1215	40	4	1088	3897	113	104	22	64	24
1040	1	67	8	0	0	0	0	8	189	1	3	32	542	3	5	1	0	42
362	1	1	1	0	3	0	1	3	57	2	0	9	162	7	1	1	1	11
639	0	1	2	0	0	0	1	4	92	11	0	10	289	37	1	0	1	25
3386	15	1927	134	2	14	0	31	139	510	38	14	449	1802	153	28	26	93	14
7104	4	2095	30	0	3	0	16	248	1525	35	5	858	5196	105	112	9	30	29
4175	1	16	10	0	0	0	0	571	952	11	2	1900	2924	39	15	3	0	13
7917	3	74	27	0	8	0	9	595	1759	65	4	2098	5419	238	30	10	26	32
9736	7	22	12	0	4	0	17	570	2014	46	0	2033	6581	154	59	20	50	36
1183	1	1	18	0	1	0	6	17	129	2	2	63	385	5	6	4	10	32
1937	32	5	78	2	85	9	19	45	317	6	130	144	957	20	35	108	60	72
7450	78	54	136	1	107	3	60	199	1447	27	12	687	4621	77	73	254	158	29
1734	0	353	163	1	107	1	10	10	326	28	26	41	1129	84	7	0	22	66

DOMICILIOS_PRECARIOS	COABITACAO	ONUS_EXCESSIVO	ADENSAMENTO	DEFICIT_TOTAL
6	54	71	0	131
23	269	472	42	807
6	136	102	20	264
0	41	29	0	70
14	44	34	0	92
16	139	46	0	202
3	24	18	5	49
1	37	11	4	54
79	282	154	19	534
473	178	15	0	667
9	302	324	53	689
8	50	66	5	129
10	8	12	0	30
2	20	3	0	25
11	223	91	19	344
0	461	327	7	795
2	212	121	21	357
39	565	408	20	1032
38	545	280	24	887
0	93	16	2	111
0	143	57	5	206
90	684	265	79	1116
4	118	42	17	179
15	169	226	13	423

Banco de Dados - Áreas de Ponderação MG

2. CORRELAÇÃO LINEAR E MAPA DE CALOR

Análise visual das variáveis de entrada e impactos individuais nas saídas.



- **Correlação positiva perfeita (+1):** Quanto mais a variável de entrada aumenta, mais a saída tende a aumentar de forma perfeitamente previsível e linear;

- **Ausência de correlação linear (0.0):** Não há relação linear discernível entre as variáveis. Importante: pode ainda existir relação não-linear;

- **Correlação negativa perfeita (-1):** Quanto mais a variável de entrada aumenta, mais a saída tende a diminuir de forma perfeitamente previsível e linear.

Variáveis com correlação próxima de +1 realmente têm um poder explicativo muito forte da variação na saída. Isso é fundamental para bons modelos. Além disso tem impactos positivos nas métricas R^2, RMSE e MAE.

Múltiplas colunas próximas de +1 geram multicolinearidade, piora generalização e infla a variância.

3. MÉTRICAS DE AVALIAÇÃO PARA MODELOS DE APRENDIZADO SUPERVISIONADO

R^2, MAE, RMSE

- **R^2 (Coeficiente de Determinação)**: Mede o poder de interpretação do modelo na fase de teste, o quanto ele está indo bem na predição;

- **MAE (Erro Médio Absoluto)**: Mostra o quanto em média o modelo está errando em relação ao valor real, quanto menor o MAE melhor será a predição;

- **RMSE (Raiz do Erro Quadrático Médio)**: Mesma coisa do MAE só que elevando os valores ao quadrado antes de tirar a média, consegue mostrar bem valores discrepantes (outliers).

- **R^2 (Coeficiente de Determinação)**: Mede o poder de interpretação do modelo na fase de teste, o quanto ele está indo bem na predição;
- **MAE (Erro Médio Absoluto)**: Mostra o quanto em média o modelo está errando em relação ao valor real, quanto menor o MAE melhor será a predição;
- **RMSE (Raiz do Erro Quadrático Médio)**: Mesma coisa do MAE só que elevando os valores ao quadrado antes de tirar a média, consegue mostrar bem valores discrepantes (outliers).

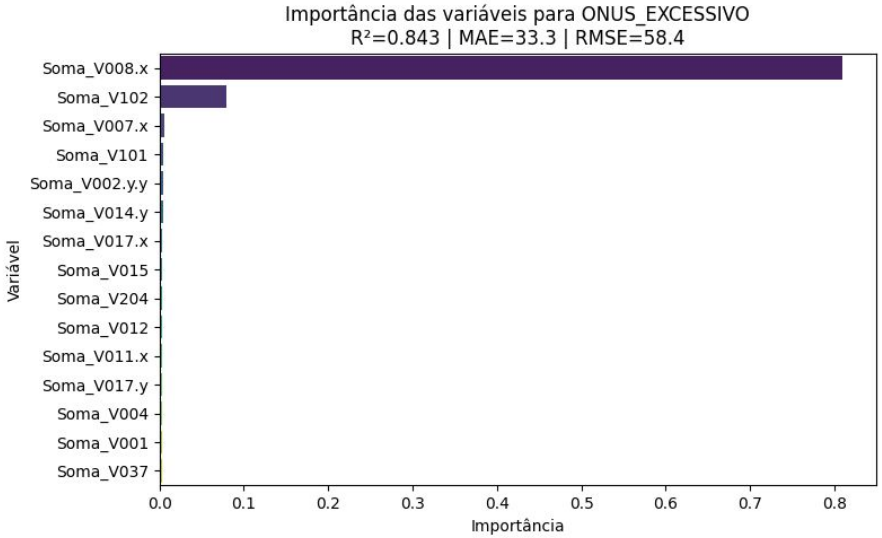
Ajustes futuros: gráficos de resíduos, cálculo do desvio padrão/variância das saídas.

https://peerj.com/articles/cs-623/

4. UTILIZANDO FLORESTA ALEATÓRIA PARA ANÁLISE DE DADOS

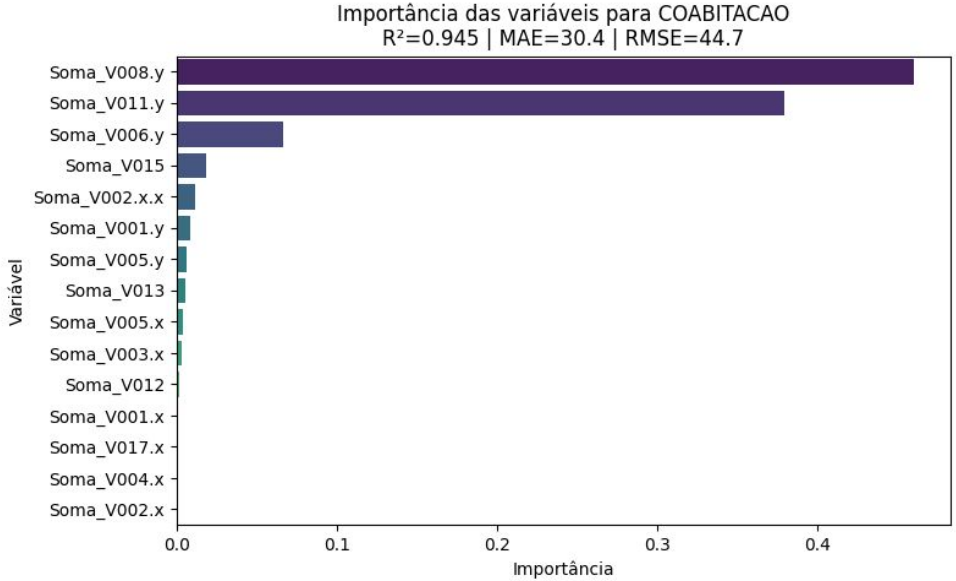
Variáveis de maior influência, análise percentual dos valores nulos (que mais contém 0's), impactos dos valores nulos individualmente em cada variável e impacto de saídas com e sem zeros para cálculo da média e desvio padrão

- **1º Etapa - Analisar individualmente as 15 variáveis mais impactantes para cada uma das 5 saídas no banco de dados de áreas de ponderação em MG utilizando 200 árvores.**



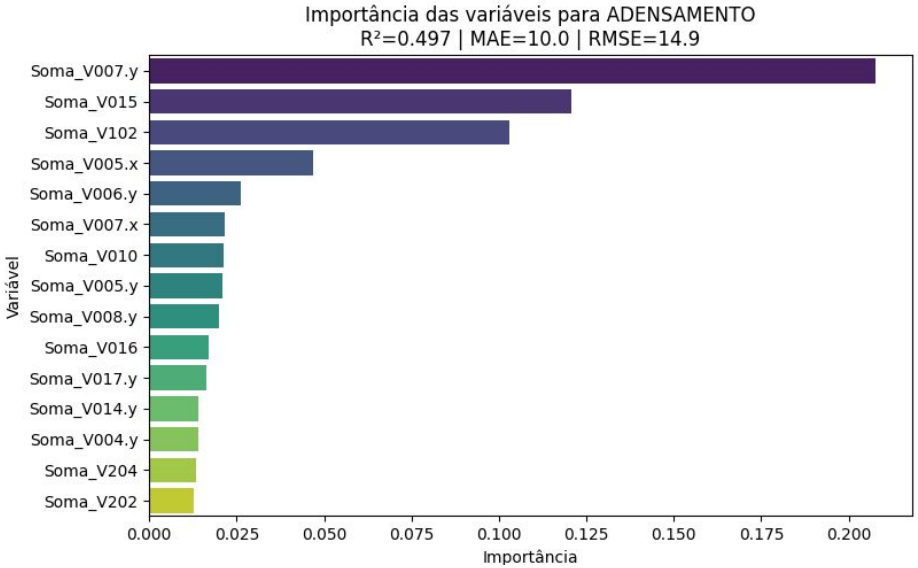
1) **ÔNUS EXCESSIVO:** O componente é quase todo zeros, e **Soma_V008.x** é a única variável com valores reais, logo o modelo só pode se apoiar nela; o R^2 alto reflete isso. Mas requer investigação no dicionário do Censo para confirmar.

- **1º Etapa - Analisar individualmente as 15 variáveis mais impactantes para cada uma das 5 saídas no banco de dados de áreas de ponderação em MG utilizando 200 árvores.**



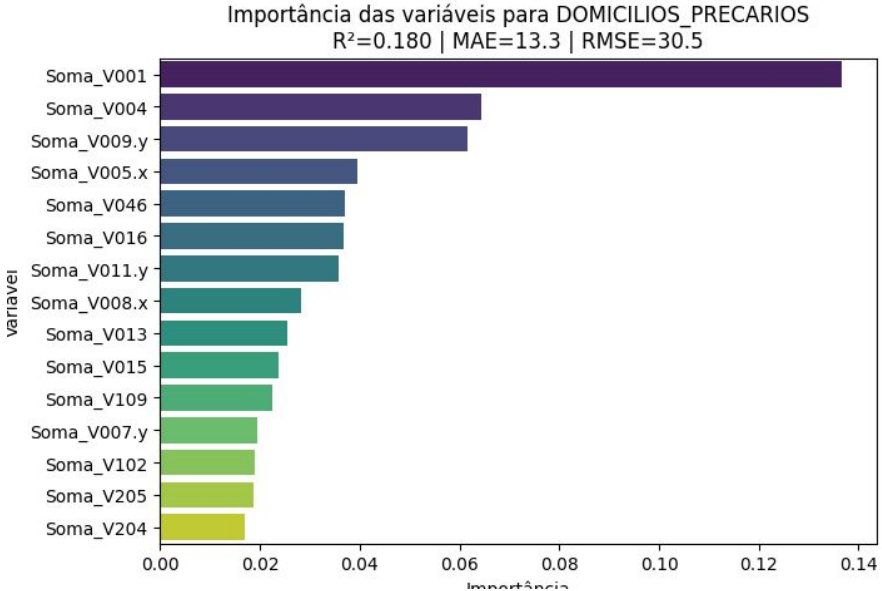
2) COABITAÇÃO: Duas variáveis dominantes (**Soma_V008.y** com 44% e **Soma_V011.y** com ~=38%) explicam bem o fenômeno. O componente é quase todo zeros, e **Soma_V008.y** e **Soma_V011.y** são as únicas variáveis com valores reais, logo o modelo só pode se apoiar nelas; o R^2 alto reflete isso. Mas requer investigação no dicionário do Censo para confirmar.

- **1º Etapa - Analisar individualmente as 15 variáveis mais impactantes para cada uma das 5 saídas no banco de dados de áreas de ponderação em MG utilizando 200 árvores.**



3) ADENSAMENTO: Performance moderada (R^2=0.497) com importância distribuída entre várias variáveis (Soma_V007.y lidera com 21%). O modelo captura parte dos padrões, mas faltam features importantes para melhor precisão.

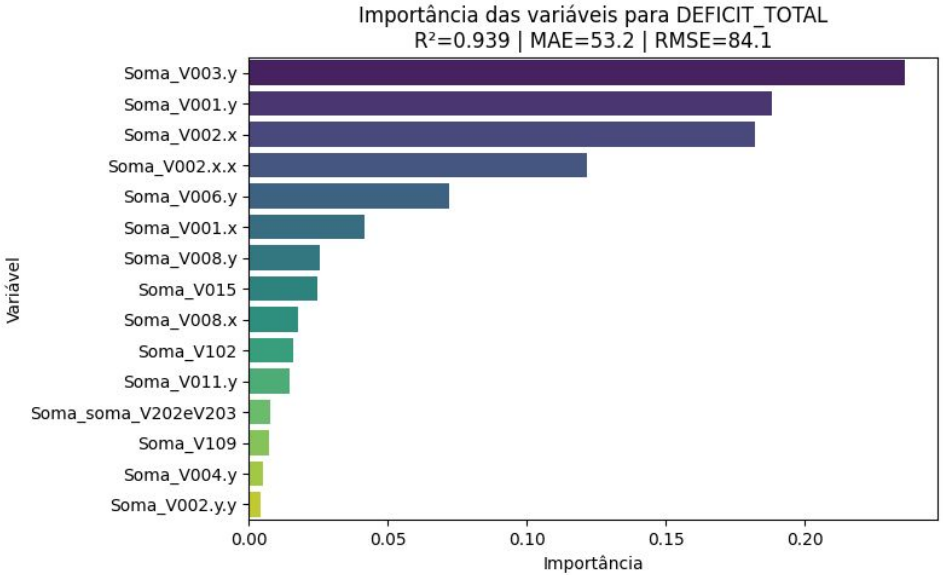
- **1º Etapa - Analisar individualmente as 15 variáveis mais impactantes para cada uma das 5 saídas no banco de dados de áreas de ponderação em MG utilizando 200 árvores.**



4) DOMICÍLIOS PRECÁRIOS: Péssimo desempenho (R^2=0.180) causado pela alta concentração de zeros na base, poucos valores moderados e alguns outliers extremos. A importância muito dispersa (nenhuma variável domina) indica que o modelo não consegue identificar padrões claros. Este é um fenômeno genuinamente difícil de prever com as variáveis disponíveis.

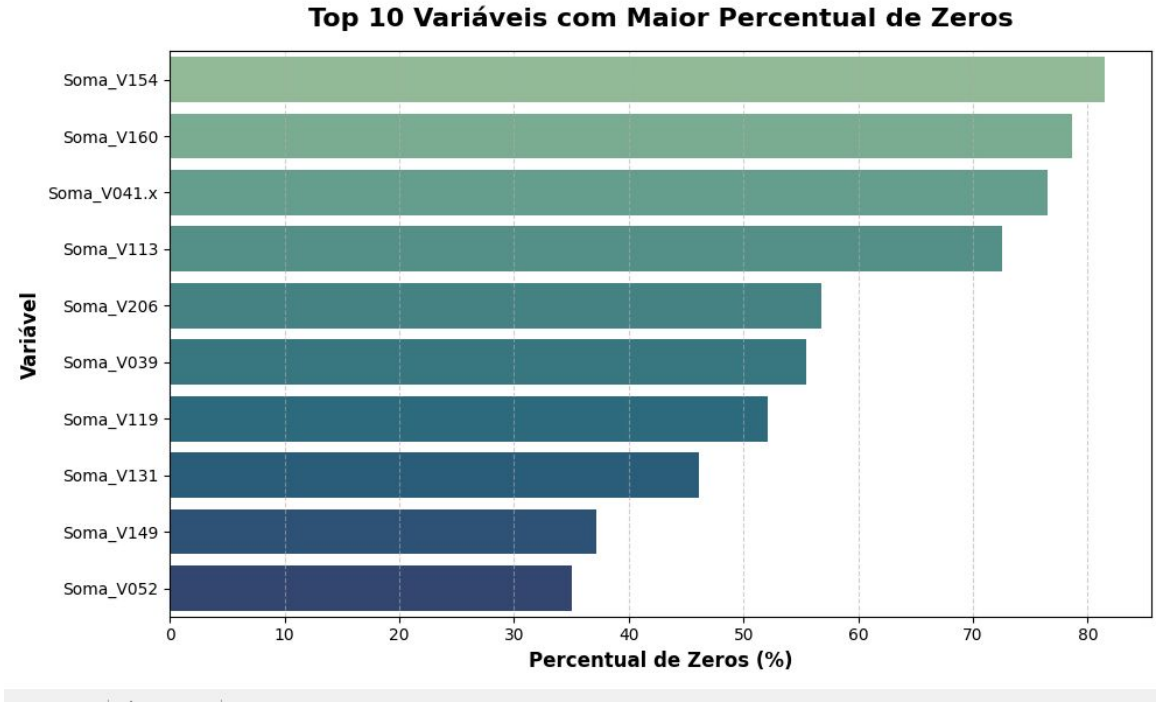
Provavelmente nenhum modelo será tão bom para esse modelo.
Sugestão: modelos de contagem.

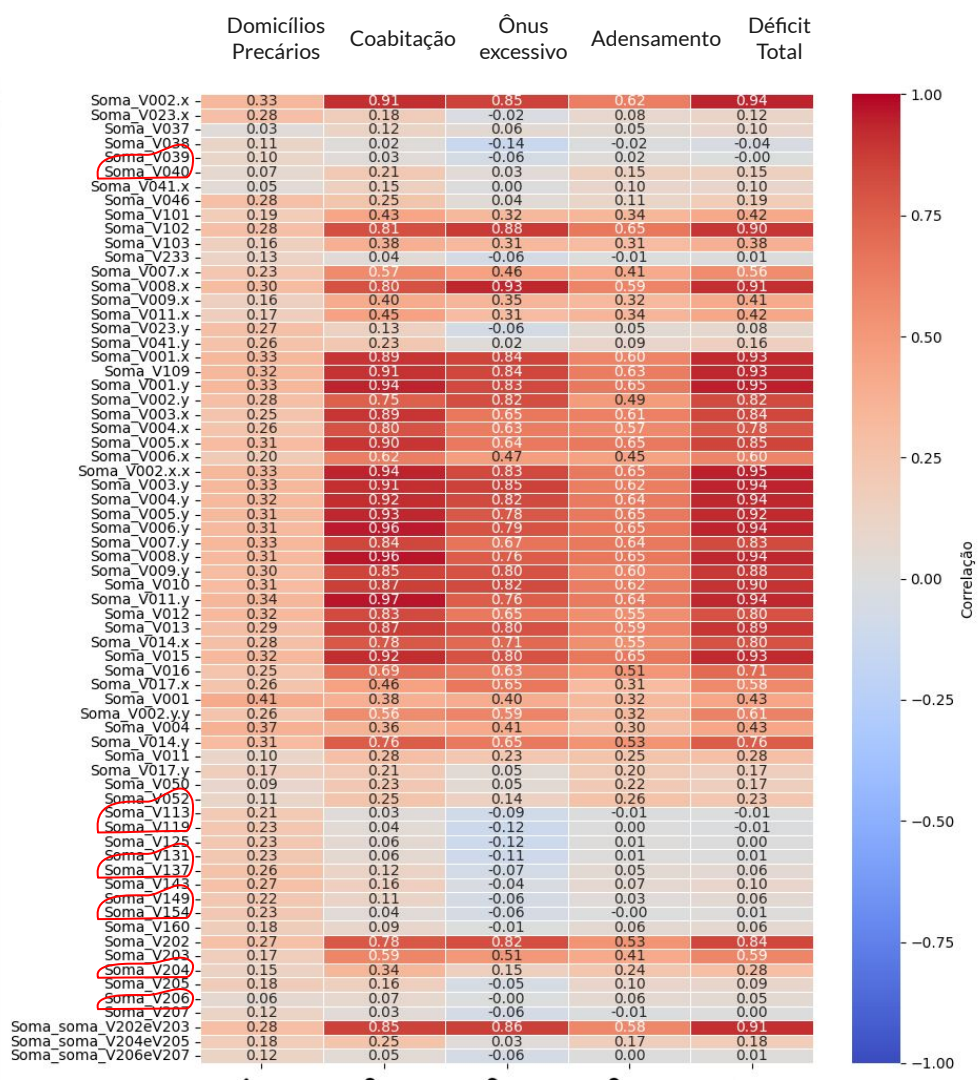
- **1º Etapa - Analisar individualmente as 15 variáveis mais impactantes para cada uma das 5 saídas no banco de dados de áreas de ponderação em MG utilizando 200 árvores.**



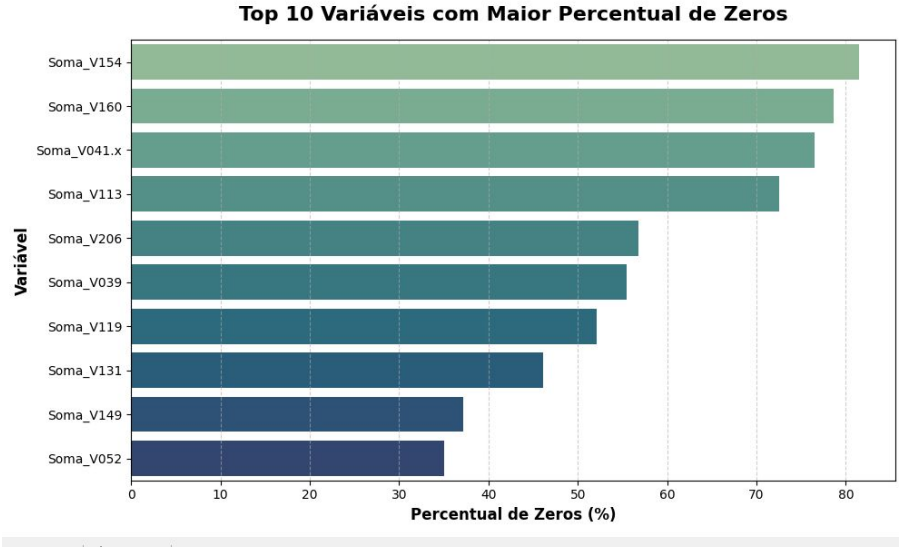
4) DÉFICIT TOTAL: Ótimo desempenho do coeficiente de rendimento(R^2 = 0.939. O modelo erra pouco em relação a unidades de déficit para a saída total. O RMSE não pune outliers discrepantes.
Conclusão: Ótimo aprendizado.

- **2º Etapa - Analisar as 10 variáveis com maior percentual de zeros em relação às suas instâncias totais no banco de dados de áreas de ponderação em MG.**





CONCLUSÃO: Variáveis com muitos zeros (~70-80%) têm correlação quase nula com as saídas, o que significa que elas são **apenas ruídos** sem informação útil. Essas variáveis devem ser **removidas do dataset** antes do treinamento, pois só atrapalham: consomem recursos computacionais, aumentam risco de overfitting e diluem a capacidade do modelo de focar nas variáveis realmente importantes. Ao fazer essa **limpeza (feature selection)**, nossos modelos (Random Forest e Rede Neural) treinarão mais rápido e provavelmente terão **métricas melhores** (R^2 maior, MAE/RMSE menores), além de serem mais interpretáveis.

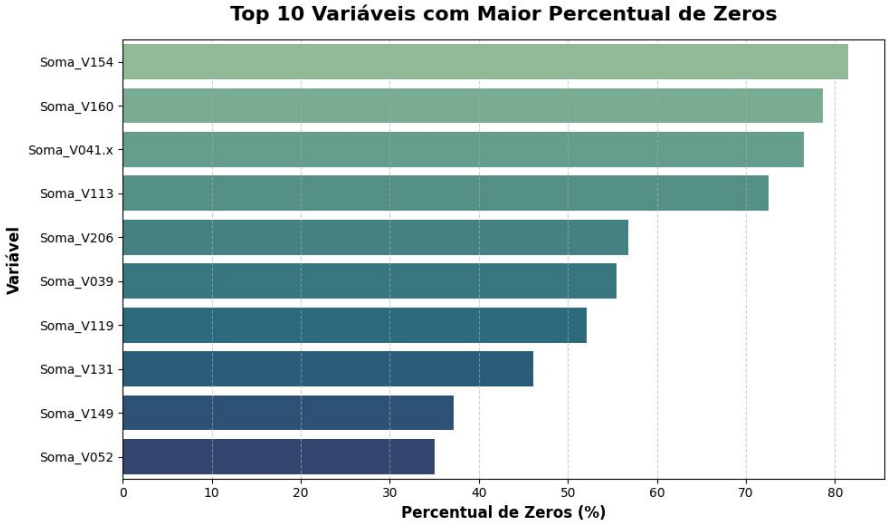


5. REAVALIANDO A ETAPA 1 SEM AS VARIÁVEIS COM MAIS ZEROS

Visualização de métricas.

- **3º Etapa - Refazer a etapa 1: “Analisar individualmente as 15 variáveis mais impactantes para cada uma das 5 saídas no banco de dados de áreas de ponderação em MG utilizando 200 árvores.” removendo as variáveis com maior percentual de zero (hipoteticamente menor impacto)**

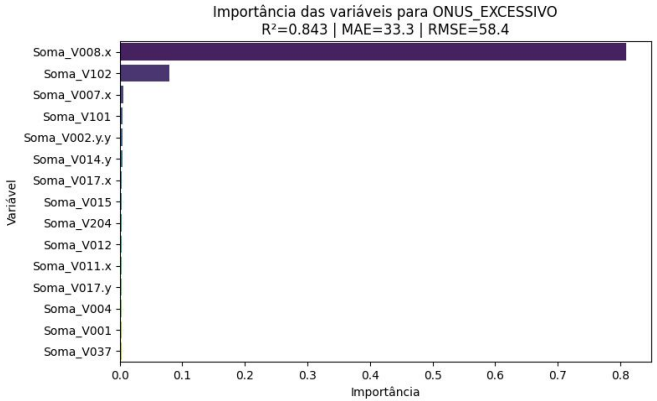
- **3º Etapa - Refazer a etapa 1: “Analisar individualmente as 15 variáveis mais impactantes para cada uma das 5 saídas no banco de dados de áreas de ponderação em MG utilizando 200 árvores.” removendo as variáveis com maior percentual de zero (hipoteticamente menor impacto)**



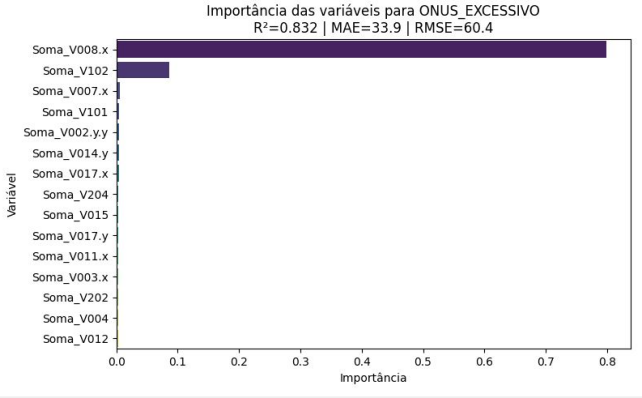

```
saidas = ["DOMICILIOS_PRECARIOS", "COABITACAO", "ONUS_EXCESSIVO", "ADENSAMENTO", "DEFICIT_TOTAL"]
dados_indesejados = ["Soma_V154", "Soma_V160", "Soma_V041.x", "Soma_V113", "Soma_V206",
                     "Soma_V039", "Soma_V119", "Soma_V131", "Soma_V149", "Soma_V052"]

entradas = [col for col in dados.columns if col not in saidas + dados_indesejados]
X = dados[entradas]
```

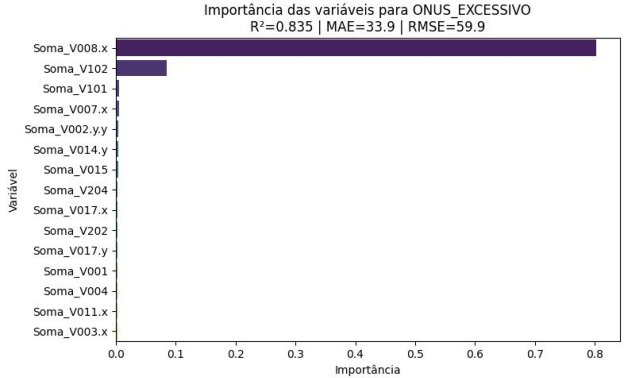
- **3º Etapa - Ônus excessivo de aluguel**

➔ **Com todas as variáveis**



➔ **Sem as 5 variáveis com maior percentual de zeros**

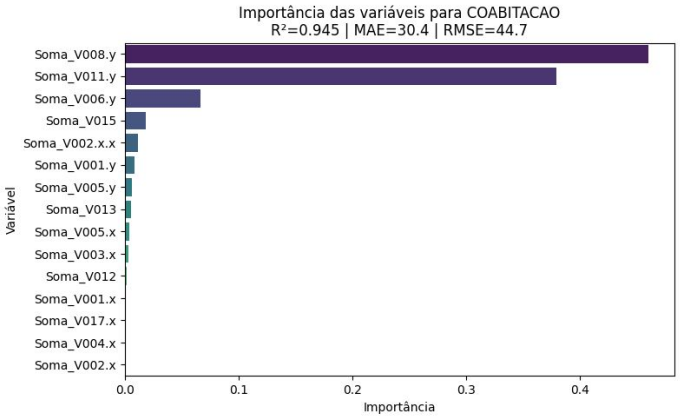




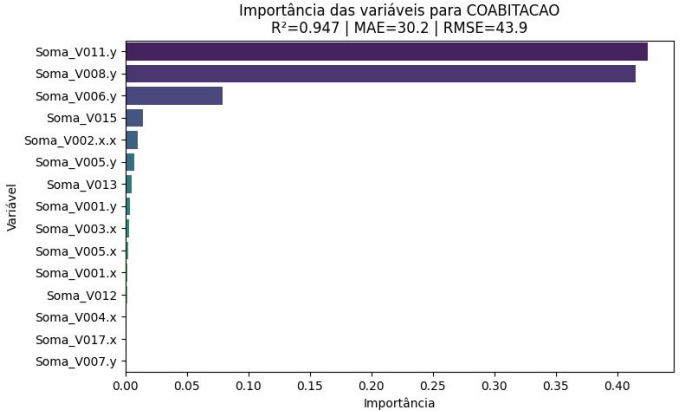
➔ **Sem as 10 variáveis com maior percentual de zeros**

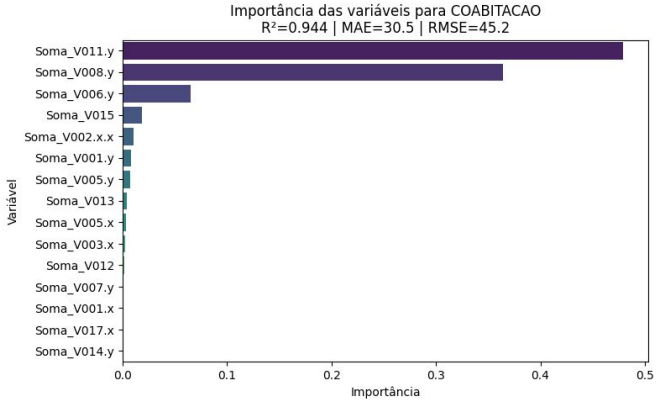
- **3º Etapa - Coabitação**

➔ **Com todas as variáveis**



➔ **Sem as 5 variáveis com maior percentual de zeros**

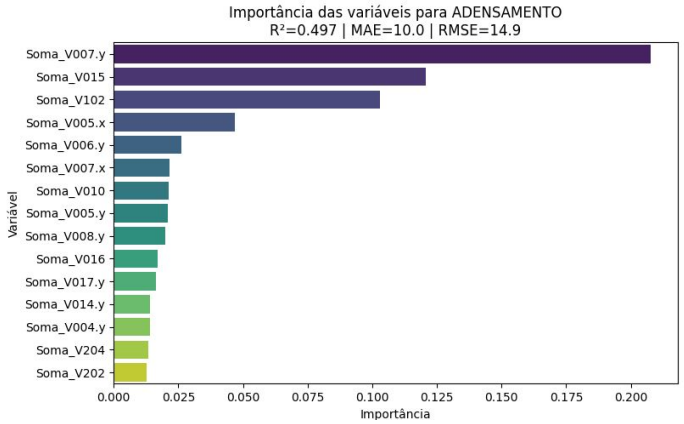




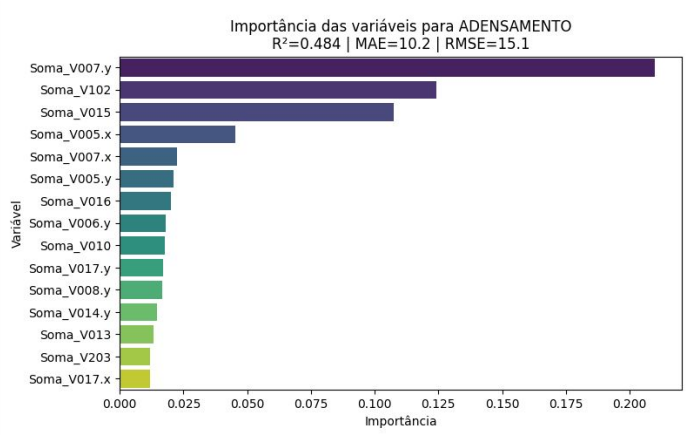
➔ **Sem as 10 variáveis com maior percentual de zeros**

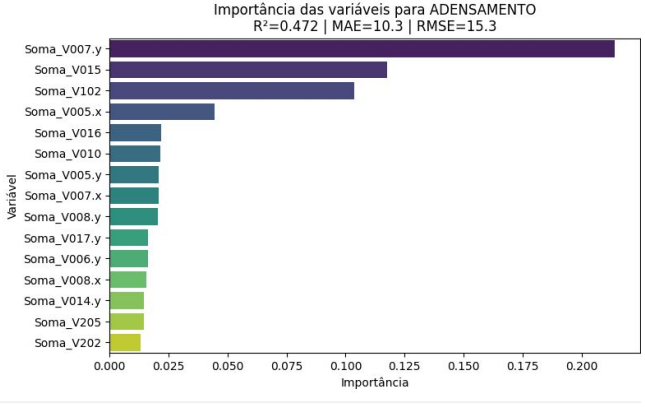
- **3º Etapa - Adensamento**

➔ **Com todas as variáveis**



➔ **Sem as 5 variáveis com maior percentual de zeros**

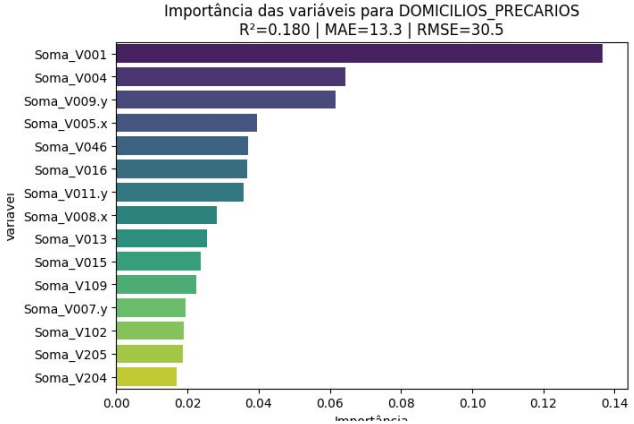




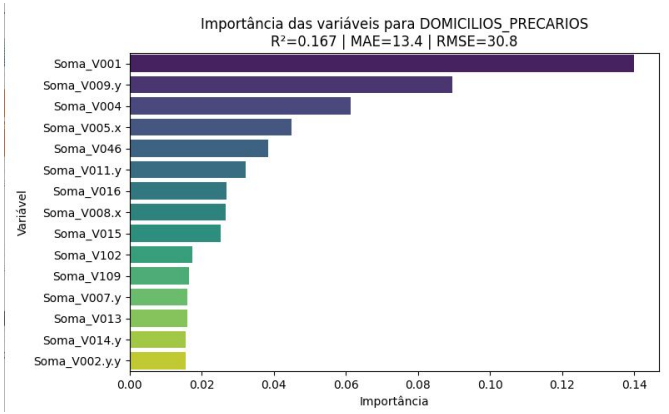
➔ **Sem as 10 variáveis com maior percentual de zeros**

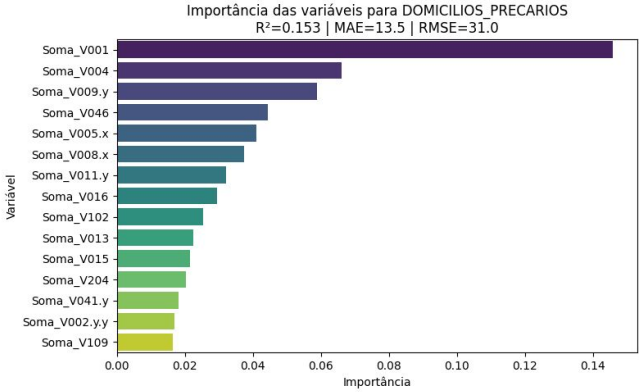
- **3º Etapa - Domicílios precários**

➔ **Com todas as variáveis**



➔ **Sem as 5 variáveis com maior percentual de zeros**

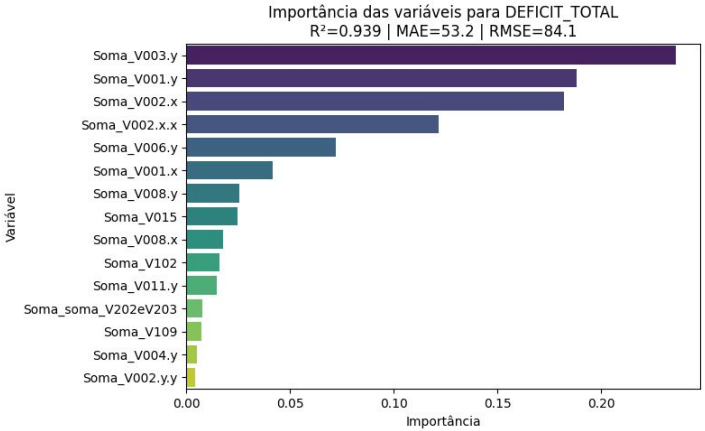




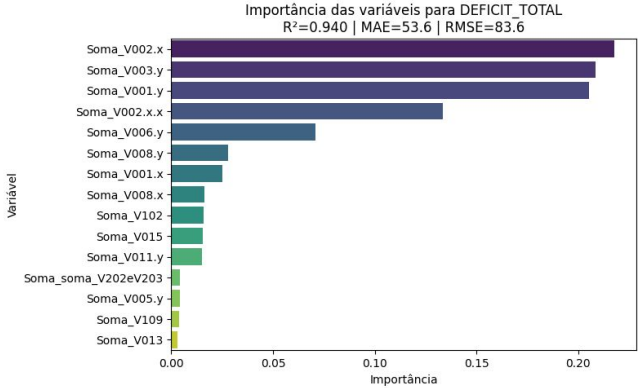
➔ **Sem as 10 variáveis com maior percentual de zeros**

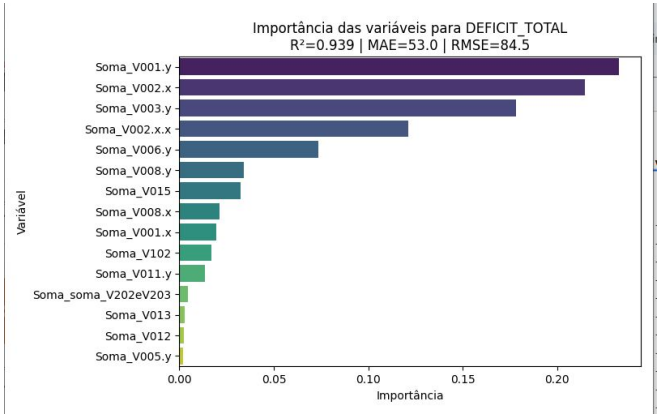
- **3º Etapa - Déficit total**

➔ **Com todas as variáveis**



➔ **Sem as 5 variáveis com maior percentual de zeros**





➔ **Sem as 10 variáveis com maior percentual de zeros**