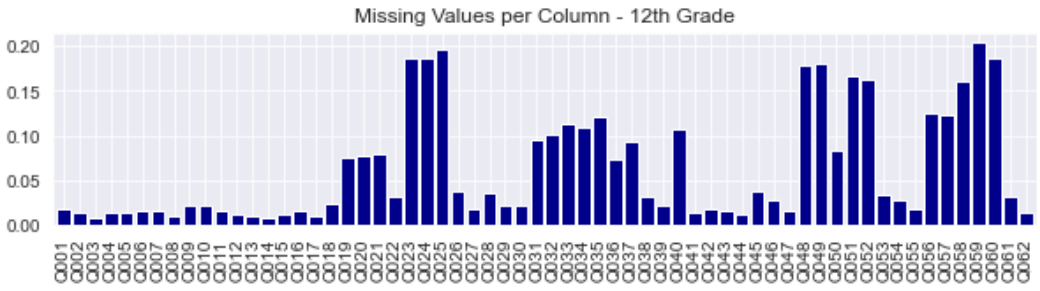
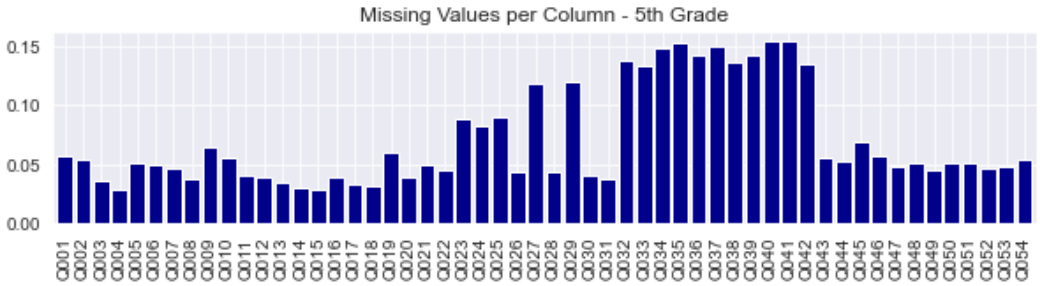
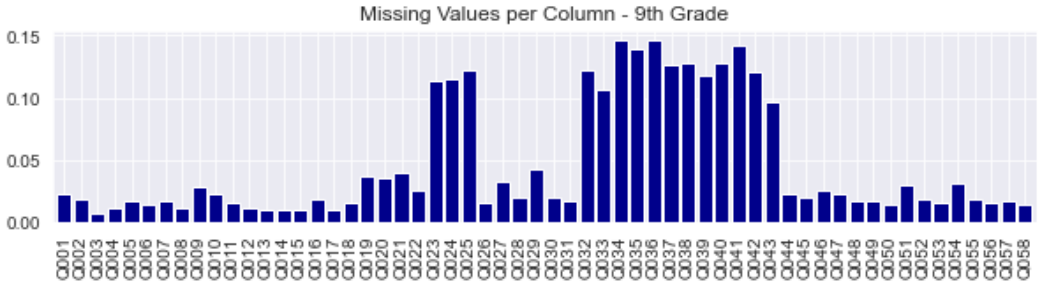
Data Cleaning



Data Cleaning - Missing Values per Column - 12th Grade



Data Cleaning - Missing Values per Column - 5th Grade



Data Cleaning - Missing Values per Column - 9th Grade

Data Cleaning

questão	TX_RESP_Q001	TX_RESP_Q002	TX_RESP_Q003	TX_RESP_Q004	TX_RESP_Q005
Enunciado	Sexo	Como você se considera?	Qual é o mês do seu aniversário?	Em que ano você nasceu?	Na sua casa tem televisão em cores?
A	Masculino.	Branco(a).	Janeiro.	1999 ou depois.	Sim, uma.
B	Feminino.	Pardo(a).	Fevereiro.	1998.	Sim, duas.
C	NaN	Preto(a).	Março	1997.	Sim, três ou mais.
D	NaN	Amarelo(a).	Abril	1996.	Não tem.
E	NaN	Indígena.	Maio	1995.	NaN
F	NaN	Não Sei.	Junho	1994.	NaN
G	NaN	NaN	Julho	1993.	NaN
H	NaN	NaN	Agosto	1992 ou antes.	NaN
I	NaN	NaN	Setembro	NaN	NaN
J	NaN	NaN	Outubro	NaN	NaN
K	NaN	NaN	Novembro	NaN	NaN
L	NaN	NaN	Dezembro	NaN	NaN

Data Cleaning - Dict_9th_Grade

ID_ALUNO	204319	204320	907687	1611630	1611631
PESO		1.2	1.2	1.2	1.2
PROFICIENCIA_LP		-0.543135	-2.067546	-1.014258	-1.528384
DESVIO_PADRAO_LP		0.335023	0.385162	0.350205	0.329361
PROFICIENCIA_LP_SAEB		220.061856395622	136.076928116311	194.106106033198	165.781175790285
DESVIO_PADRAO_LP_SAEB		18.4575436853444	21.2198698027736	19.293971119374	18.1456050651708
PROFICIENCIA_MT		-0.561132	-0.052366	-0.817919	-0.669903
DESVIO_PADRAO_MT		0.324852	0.355331	0.357318	0.313721
PROFICIENCIA_MT_SAEB		218.601407260817	247.037523357189	204.24898405636	212.521942862806
DESVIO_PADRAO_MT_SAEB		18.1567345029708	19.8602767650349	19.9713348205722	17.5345970011159

Data Cleaning - Grades_Cols_Head

ID_ALUNO	204319	204320	907687	1611630	1611631
ID_SAEB	2011	2011	2011	2011	2011
ID_REGIAO	1	1	1	1	1
ID_UF	11	11	11	11	11
ID_MUNICIPIO	1100015	1100015	1100015	1100015	1100015
ID_ESCOLA	11024682	11024682	11024682	11024682	11024682
ID_DEPENDENCIA_ADM	2	2	2	2	2
ID_LOCALIZACAO	1	1	1	1	1
ID_CAPITAL	2	2	2	2	2
ID_TURMA	52401	52401	52401	52401	52401
ID_TURNO	2	2	2	2	2
ID_SERIE	5	5	5	5	5
IN_SITUACAO_CENSO	1	1	1	1	1
IN_PREENCHIMENTO	0	1	1	1	1
IN_PROFICIENCIA	0	1	1	1	1

Data Cleaning - Grades_Ids_Table_Head

ID_ALUNO	204319	204320	907687	1611630	1611631
ID_SAEB	2011	2011	2011	2011	2011
ID_REGIAO	1	1	1	1	1
ID_UF	11	11	11	11	11
ID_MUNICIPIO	1100015	1100015	1100015	1100015	1100015
ID_ESCOLA	11024682	11024682	11024682	11024682	11024682
ID_DEPENDENCIA_ADM	2	2	2	2	2
ID_LOCALIZACAO	1	1	1	1	1
ID_CAPITAL	2	2	2	2	2
ID_TURMA	52401	52401	52401	52401	52401
ID_TURNO	2	2	2	2	2
ID_SERIE	5	5	5	5	5
IN_SITUACAO_CENSO	1	1	1	1	1
IN_PROVA_BRASIL	1	1	1	1	1
IN_PREENCHIMENTO	0	1	1	1	1
TX_RESP_Q001	.	A	B	A	A
TX_RESP_Q002	.	B	B	B	B
TX_RESP_Q003	.	F	F	E	C
TX_RESP_Q004	.	E	F	C	D
TX_RESP_Q005	.	A	B	A	A

Data Cleaning - Main_Table_Head

Data Transformation

Class	Count	Abandonment
0	326989	No, never
1	28219	Yes, once or more

Data Transformation - Class_Count_Filtered

Class	Count	Alternative
0	326976	No, never
1	28217	Yes, once at least

Data Transformation - Class_Count_Filtered_Grades

Class	Count	Abandonment
A	326989	No, never
B	21818	Yes, once
C	6401	Yes, twice or more
.	5247	Missing

Data Transformation - Class_Count_Original

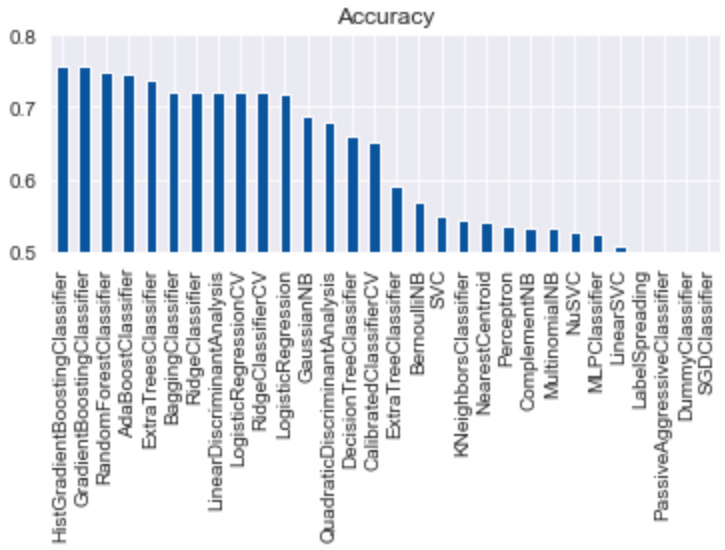
Grade	Student Count
5	468706
9	360455
12	20117
Total	849278

Data Transformation - Student_Count

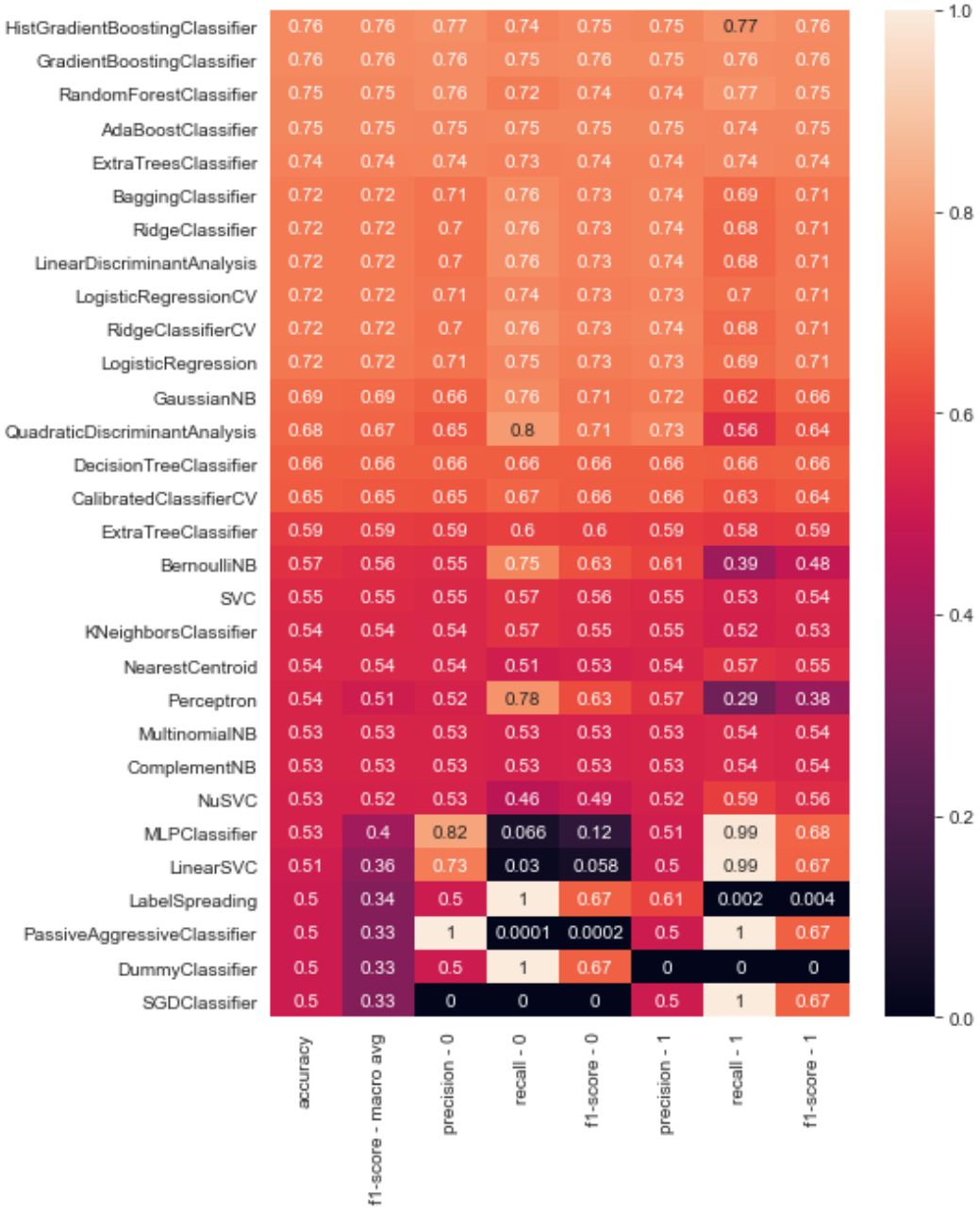
questão	TX_RESP_Q050
Enunciado	Você já abandonou a escola durante o período de aulas e ficou fora da escola o resto do ano?
A	Não.
B	Sim, uma vez.
C	Sim, duas vezes ou mais.

Data Transformation - Target_Variable_Info

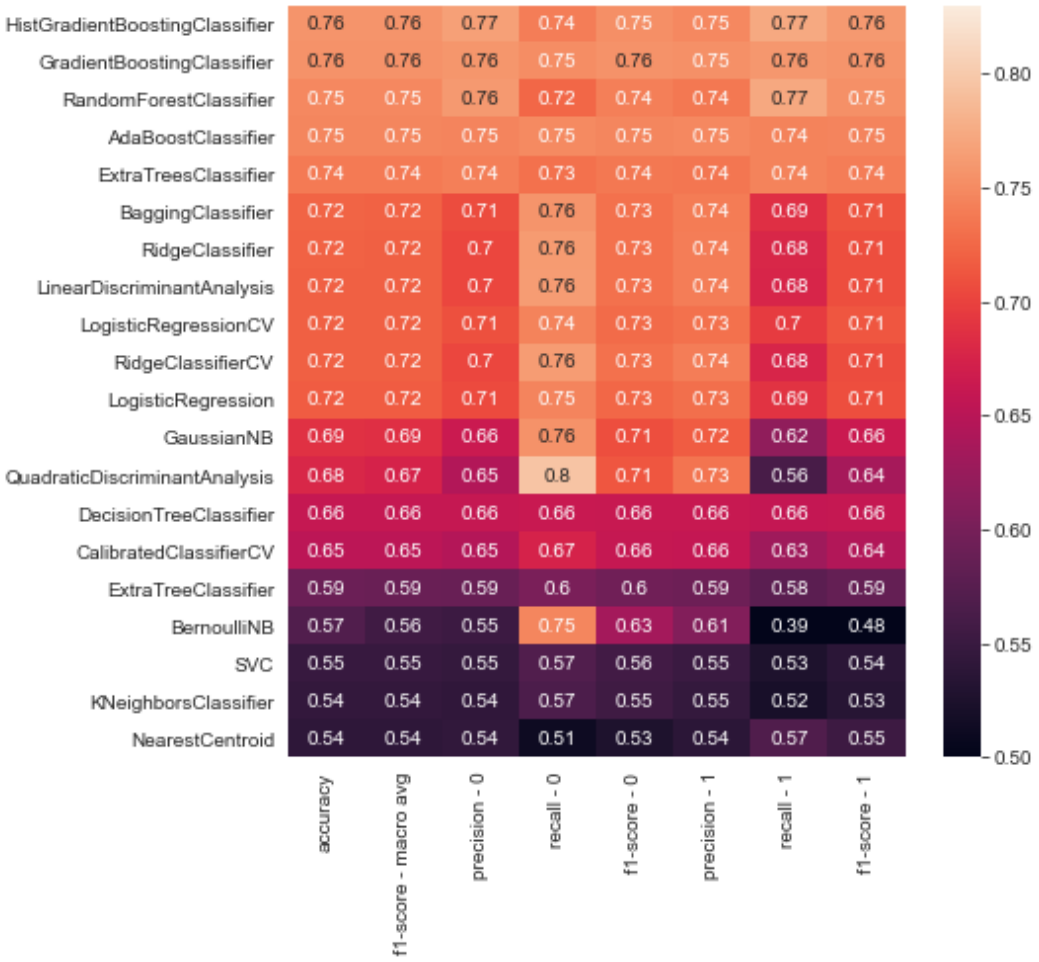
Evaluation



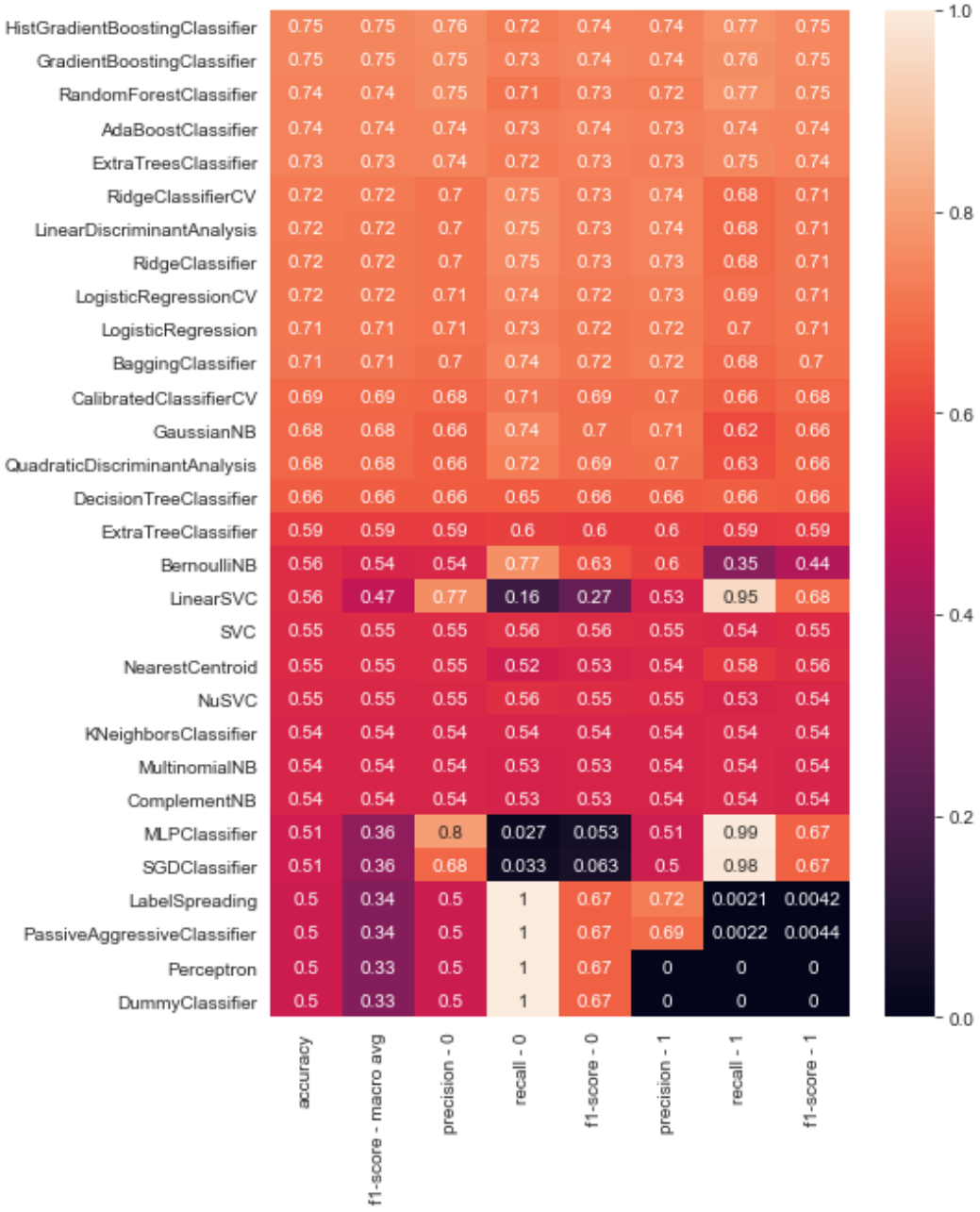
Evaluation - Classification_Comparison_Barplot



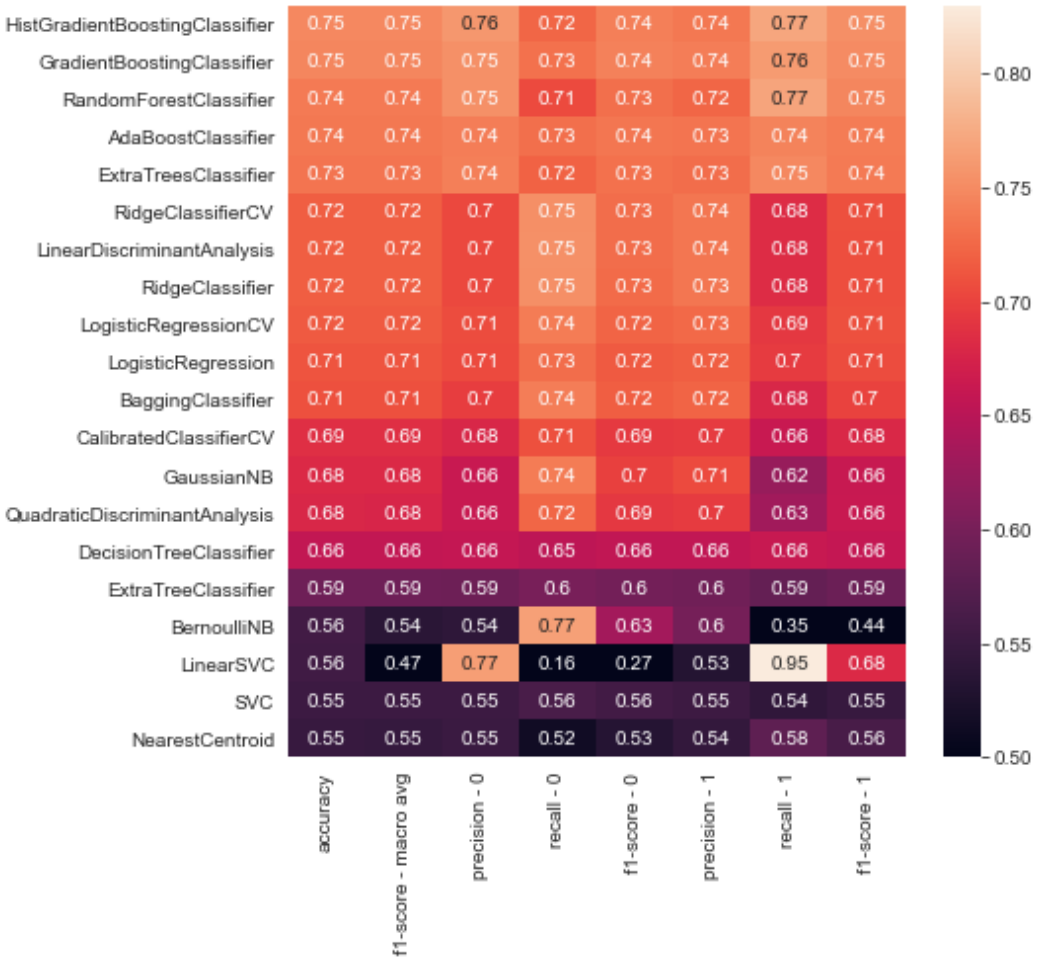
Evaluation - Classification_Comparison_Heatmap



Evaluation - Classification_Comparison_Heatmap_head



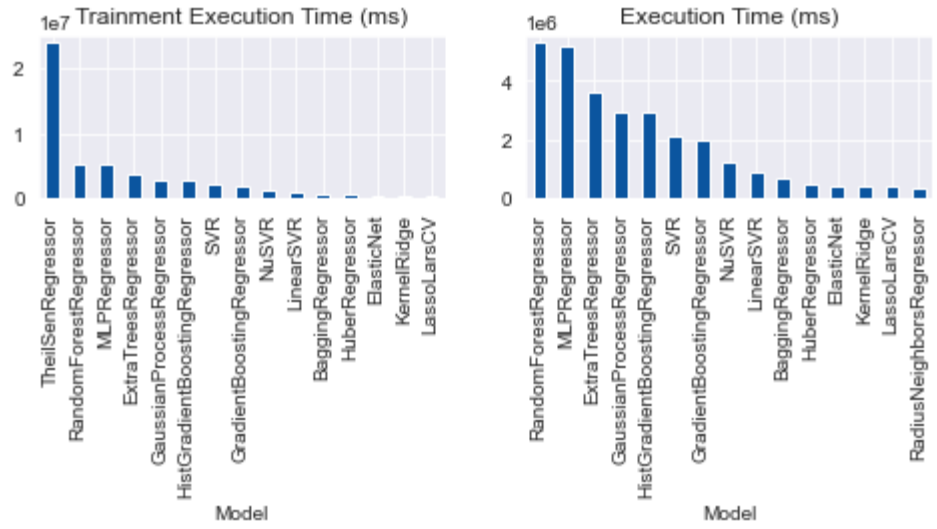
Evaluation - Model_Comparison_Heatmap



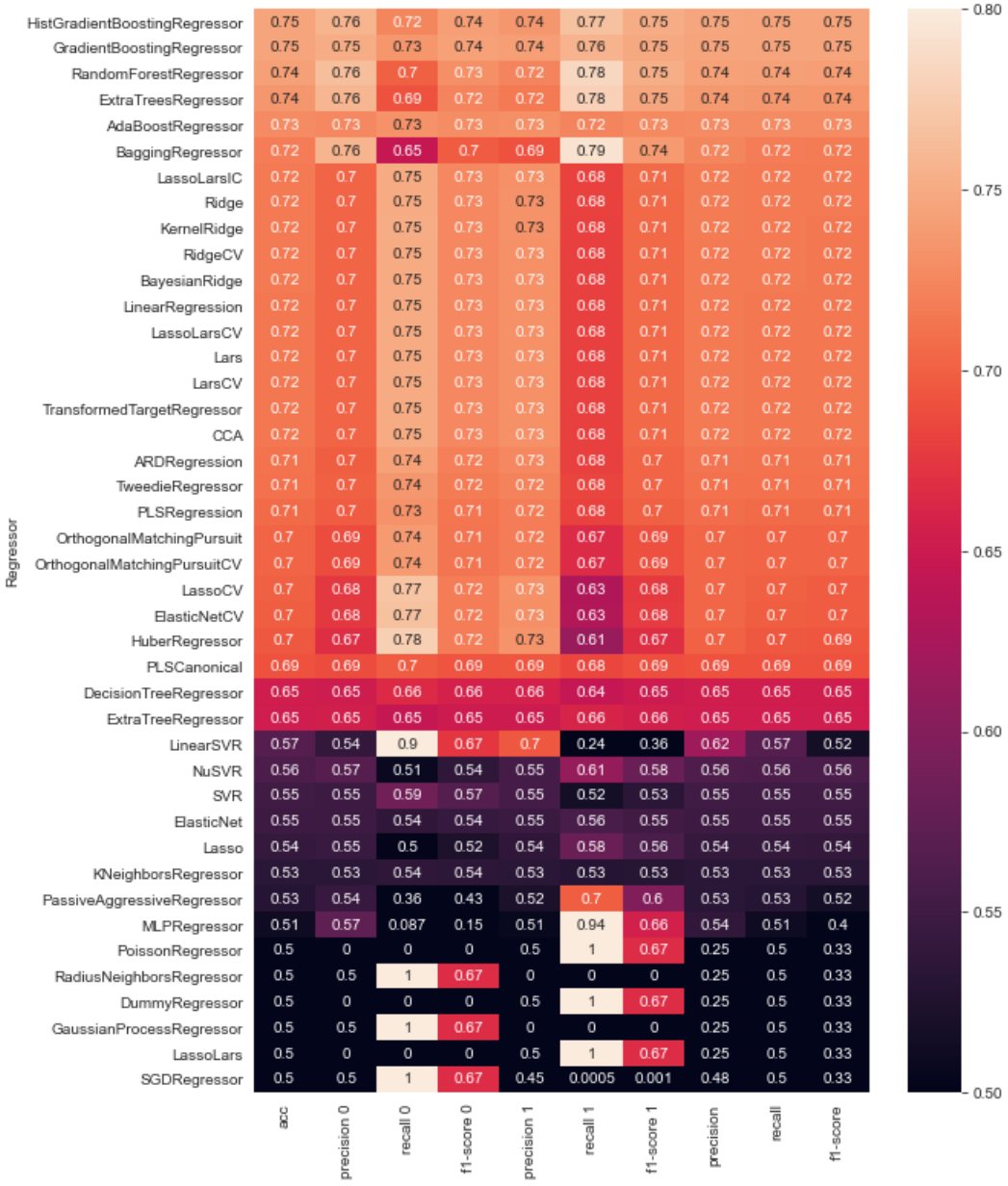
Evaluation - Model_Comparison_Heatmap_head

	Original	Treinamento	Teste
Regular	326976	6000	10000
Abandono	28217	6000	10000

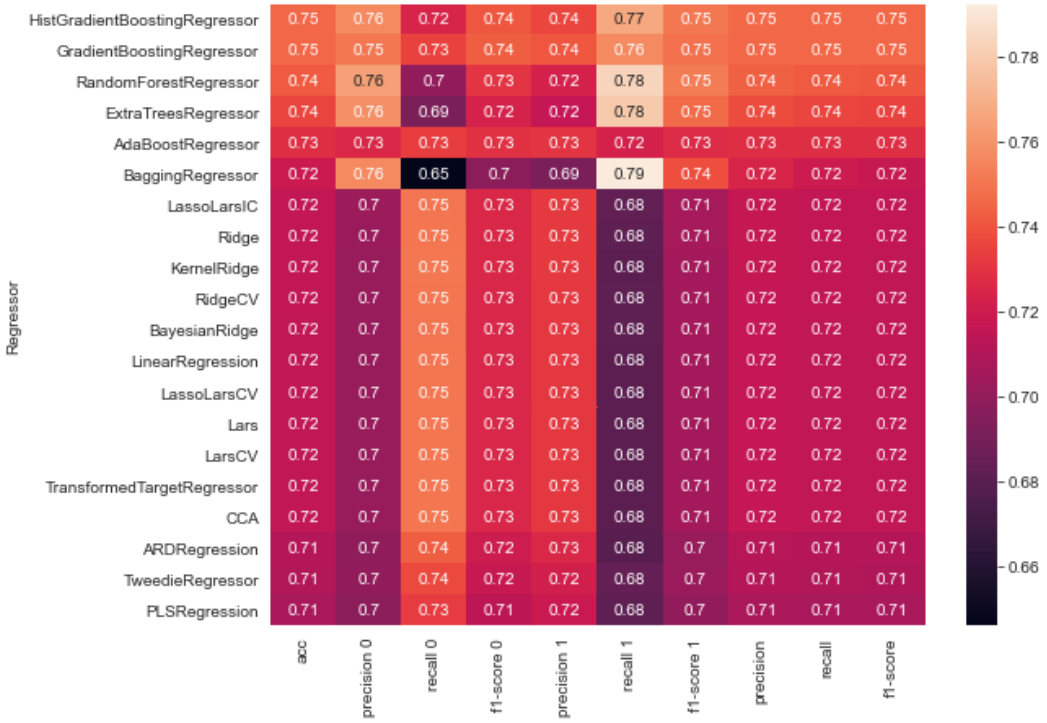
Evaluation - Model_Comparison_Sampling



Evaluation - Regression - Execution Time



Evaluation - Regression_Comparison_Heatmap

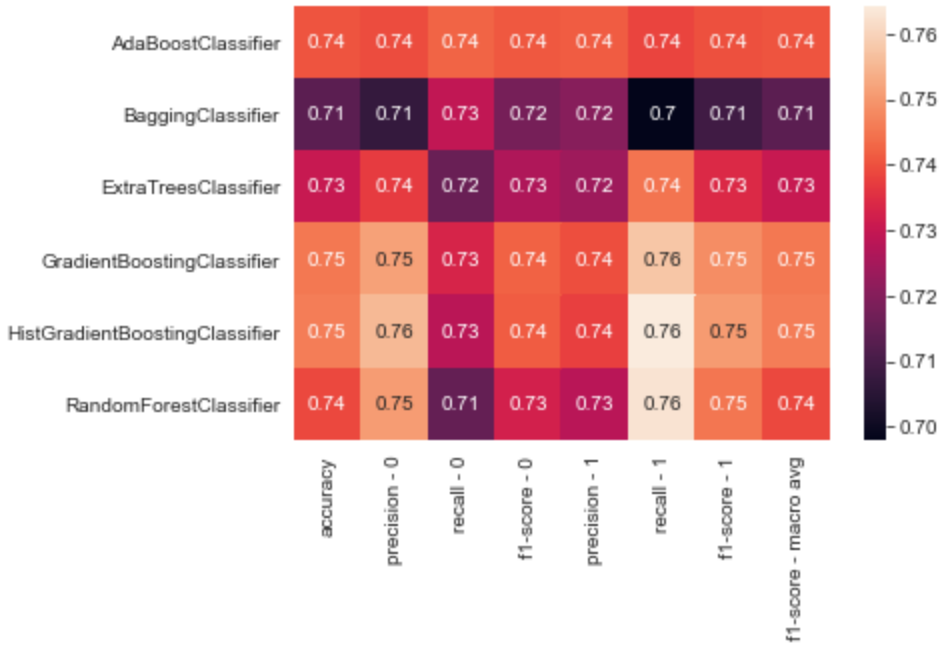


Evaluation - Regression_Comparison_Heatmap_head

Evaluation

	Original	Treinamento	Teste
Regular	326976	3000	10000
Abandono	28217	3000	10000

Evaluation - Sampling

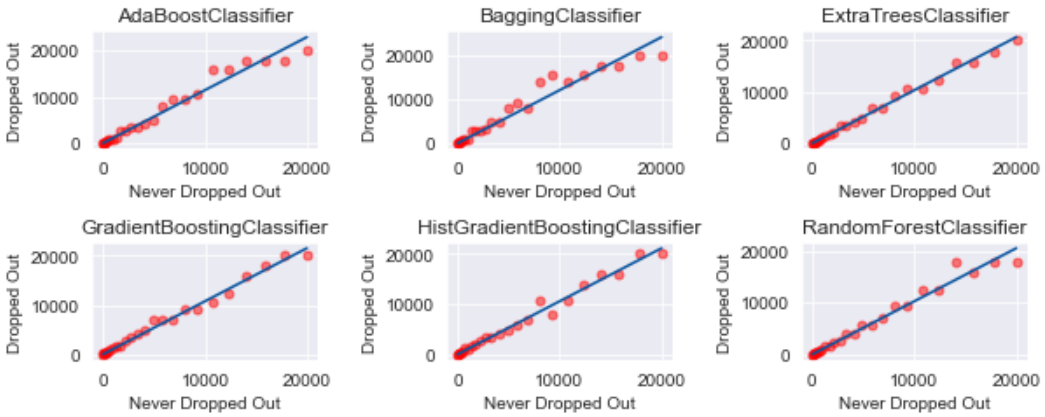


Model_Comparison_Heatmap

	Original	Treinamento	Teste
Regular	326976	23000	5000
Abandono	28217	23000	5000

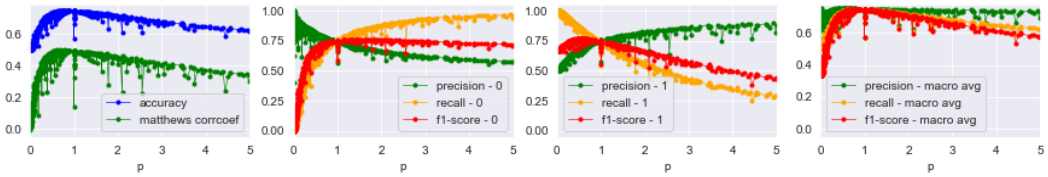
Optimization_Sampling

Class Proportion Optimization

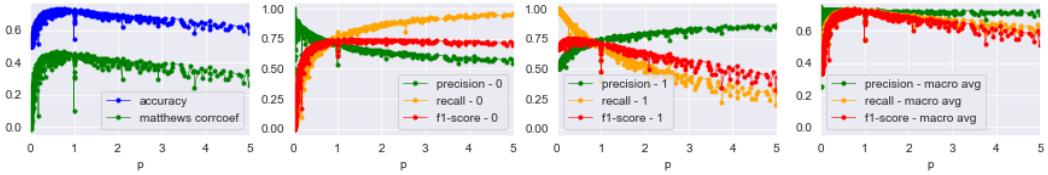


Class Proportion Optimization - 2D Regression

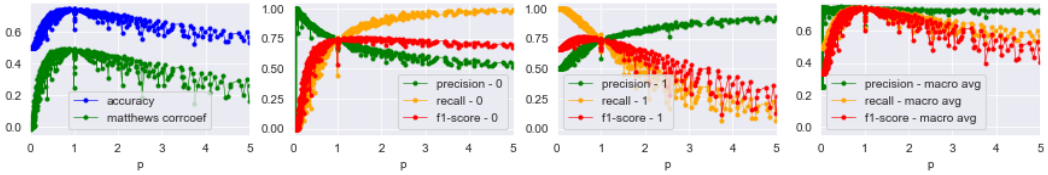
Class Proportion Optimization Line Plots



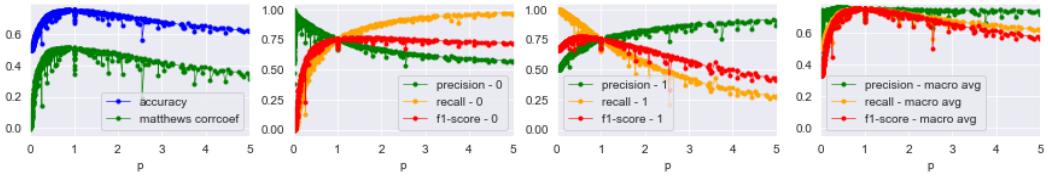
Line Plots - AdaBoostClassifier



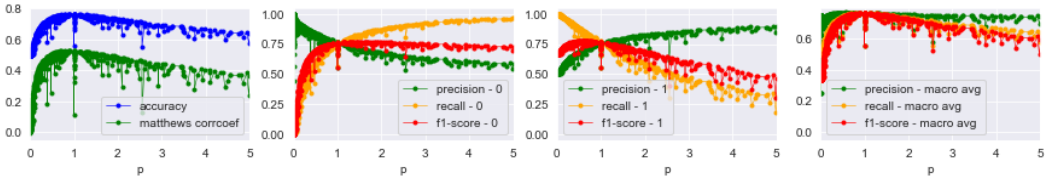
Line Plots - BaggingClassifier



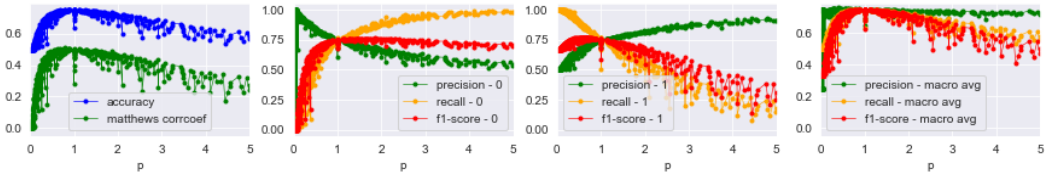
Line Plots - ExtraTreesClassifier



Line Plots - GradientBoostingClassifier



Line Plots - HistGradientBoostingClassifier



Line Plots - RandomForestClassifier

model	a	r2
AdaBoostClassifier	1.139665	0.965464
BaggingClassifier	1.213678	0.939322
ExtraTreesClassifier	1.033014	0.993009
GradientBoostingClassifier	1.076263	0.989199
HistGradientBoostingClassifier	1.057620	0.984451
RandomForestClassifier	1.026000	0.974611

Optimal Class Proportions - Regression

Class Proportion Optimization

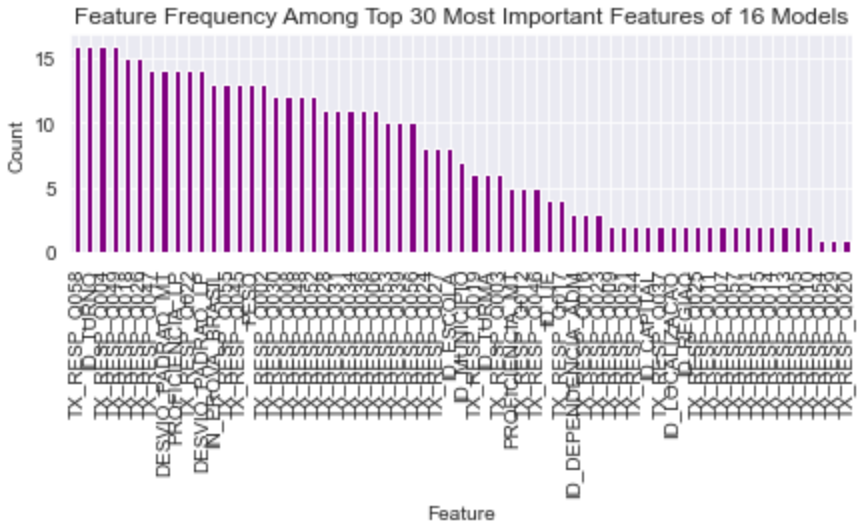
model	a	r2
AdaBoostClassifier	0.970131	0.969532
BaggingClassifier	1.232373	0.731826

Class Proportion Optimization - Optimal Class Proportions - Regression

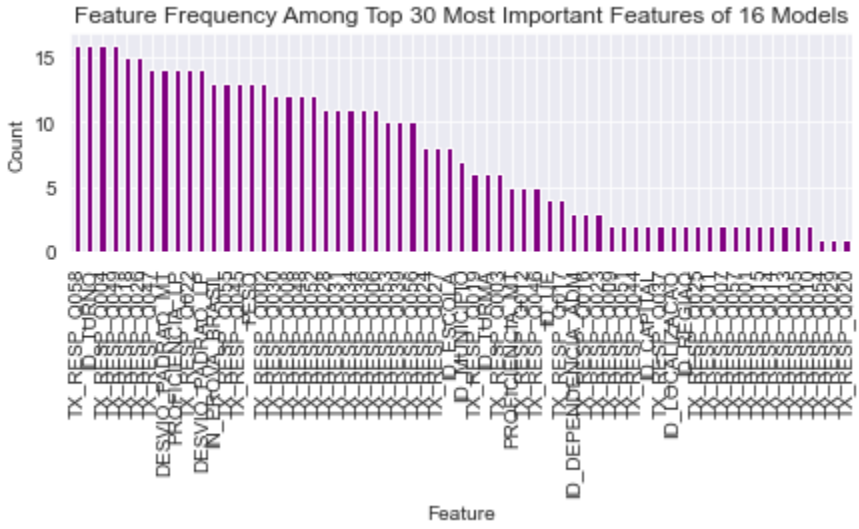
	Original	Treinamento	Teste
Regular	326976	20000	5000
Abandono	28217	20000	5000

Class Proportion Optimization - Sampling

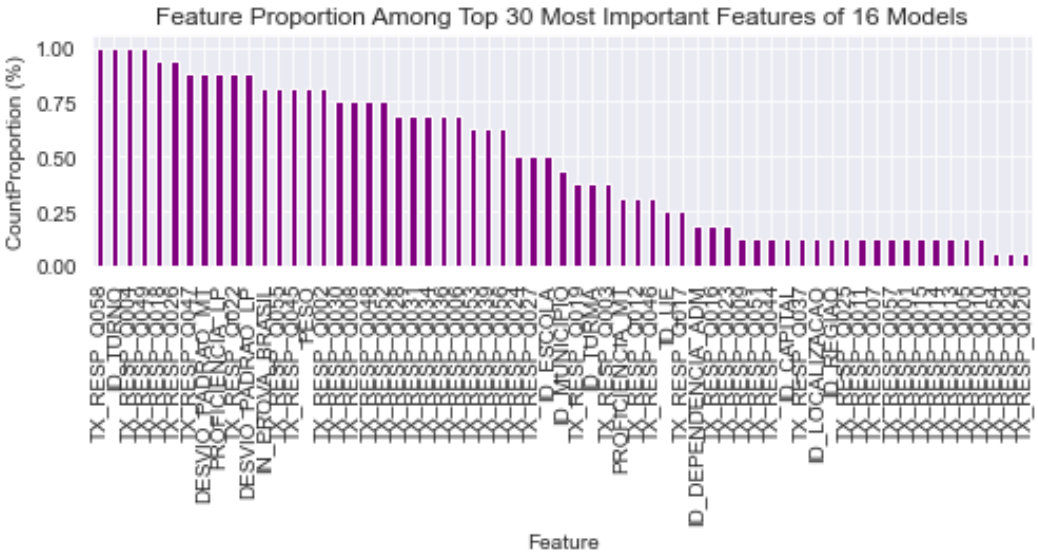
Feature Importance



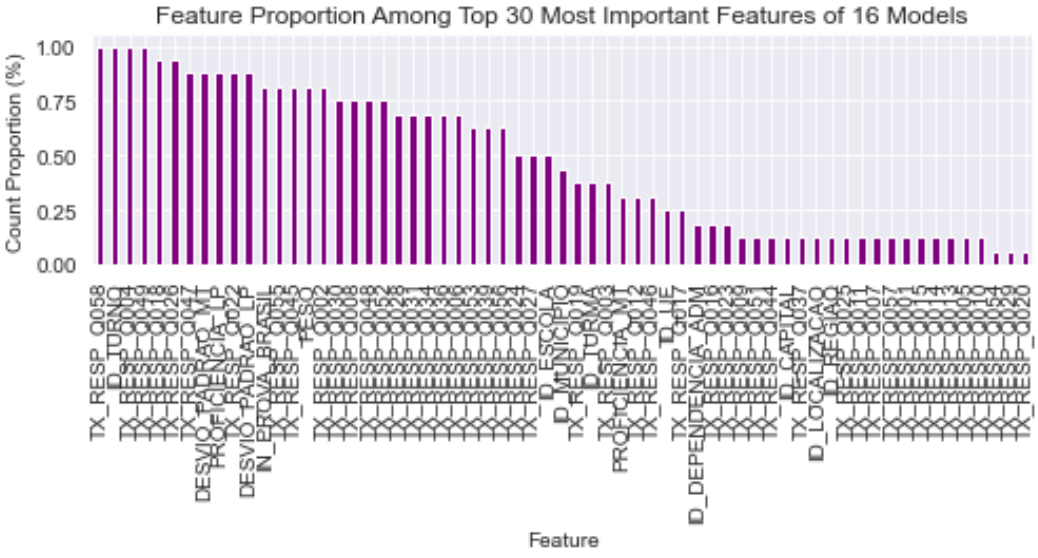
Feature Importance - Feature Frequency Barplot - Top 10



Feature Importance - Feature Frequency Barplot - Top 30

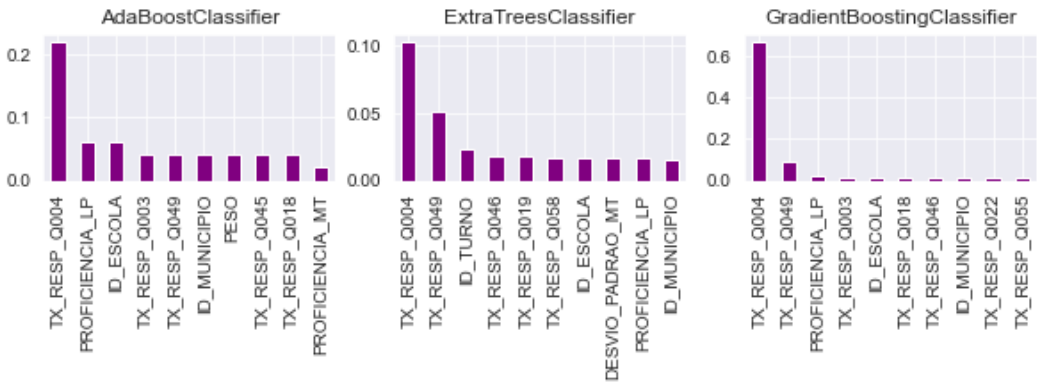


Feature Importance - Frequency Proportion Barplot - Top 10

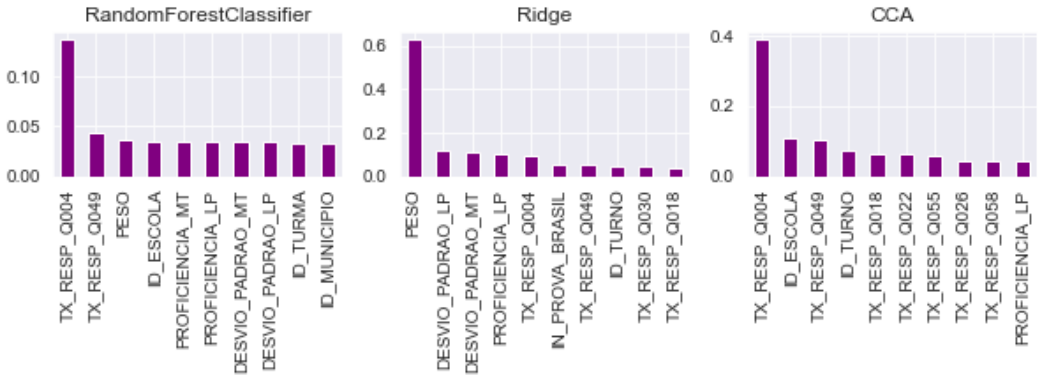


Feature Importance - Frequency Proportion Barplot - Top 30

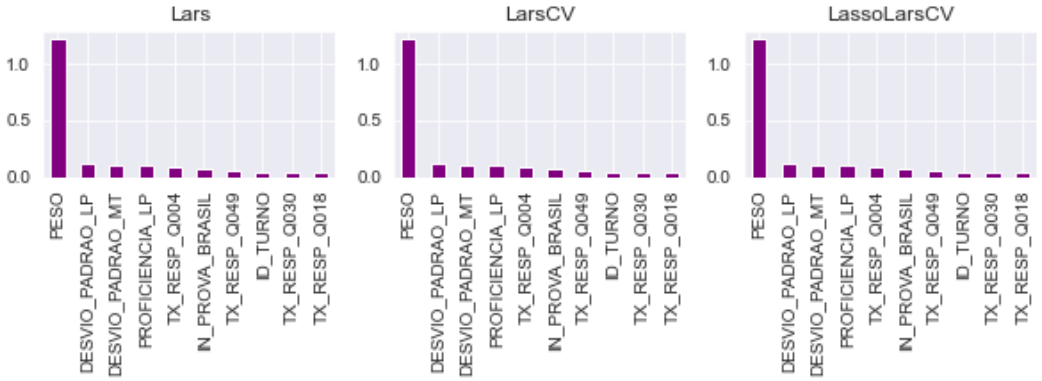
Feature Importance Per Model



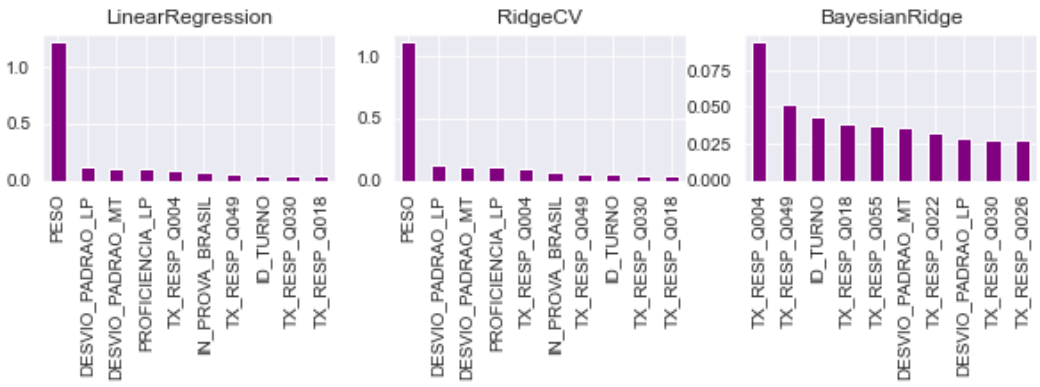
Per Model - Bar plots - 1



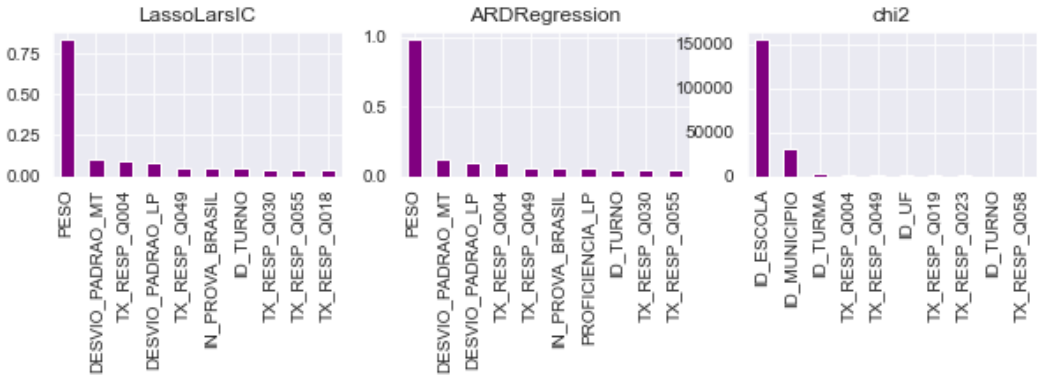
Per Model - Bar plots - 2



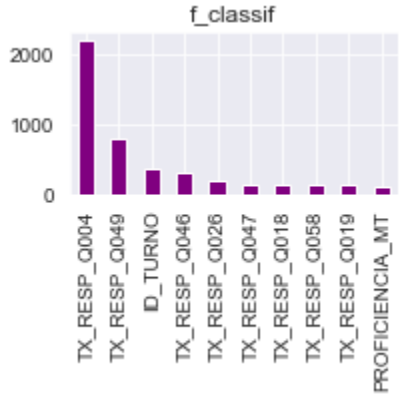
Per Model - Bar plots - 3



Per Model - Bar plots - 4

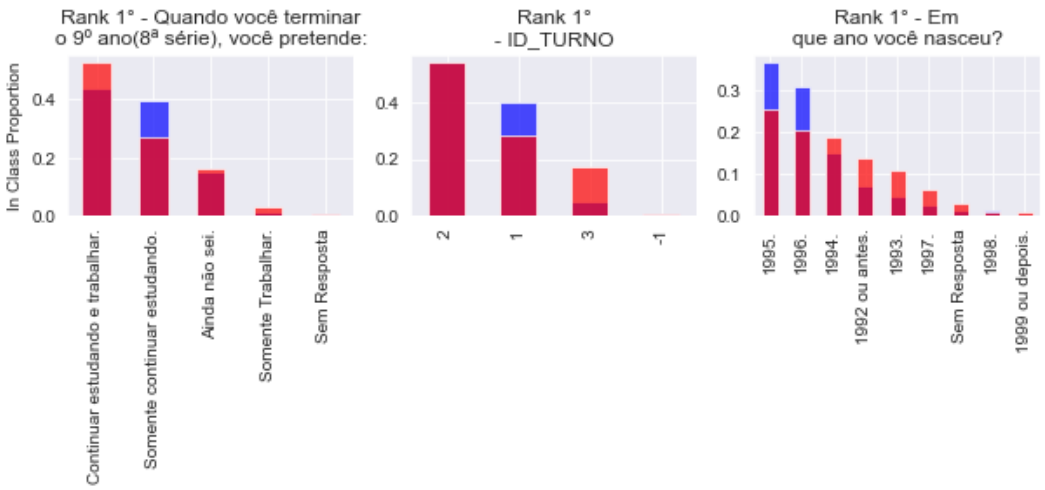


Per Model - Bar plots - 5

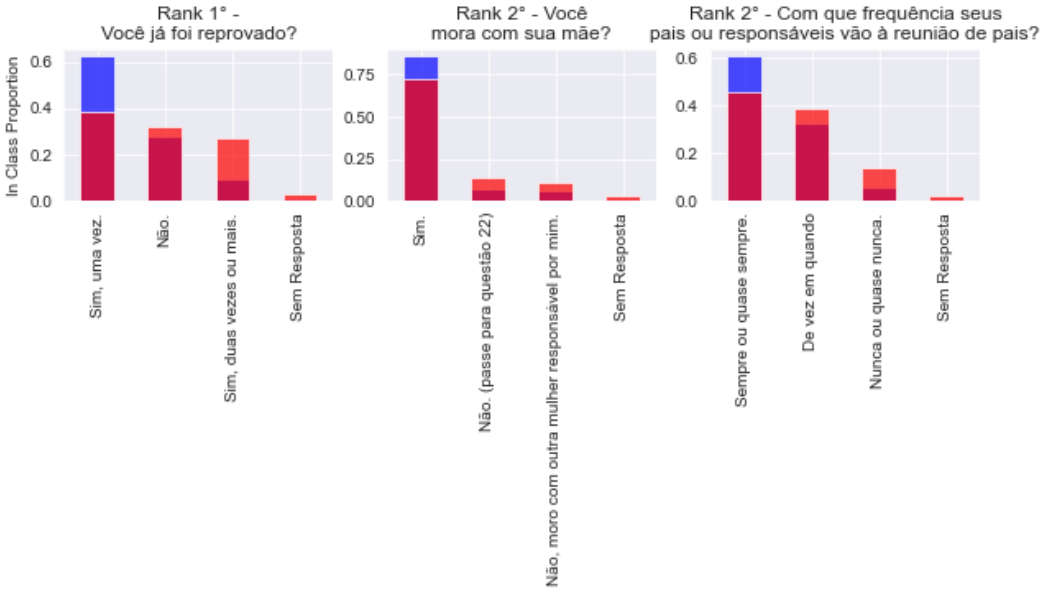


Per Model - Bar plots - 6

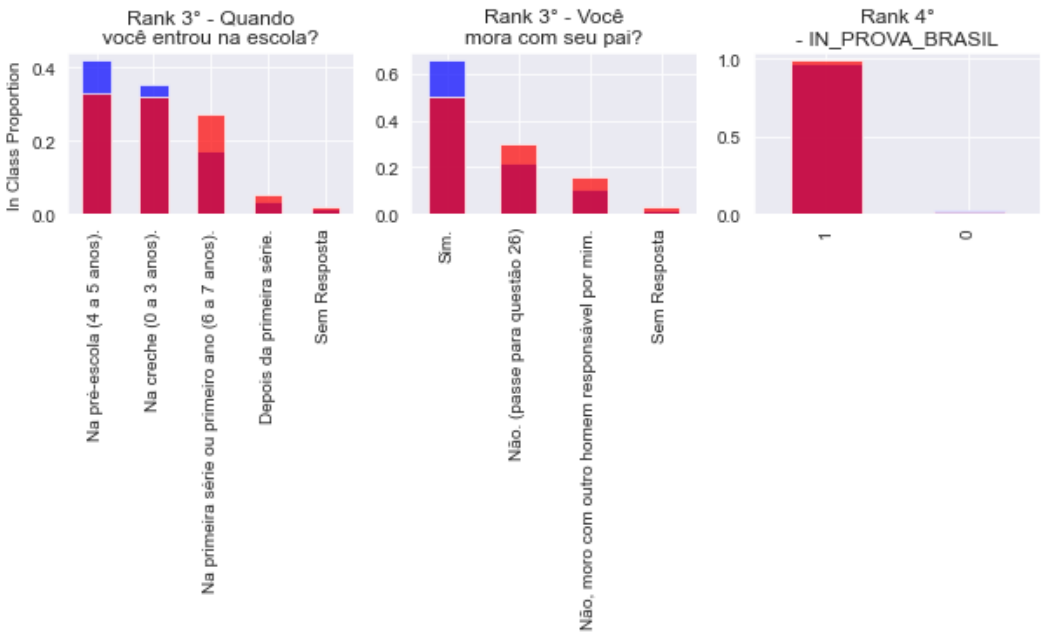
Feature Importance Feature Proportion



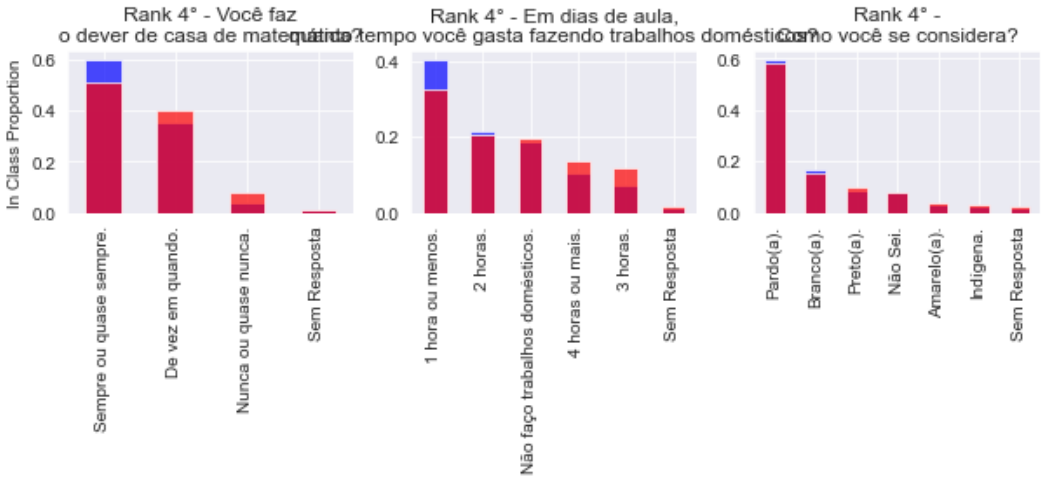
Feature Proportion - In Class - 1



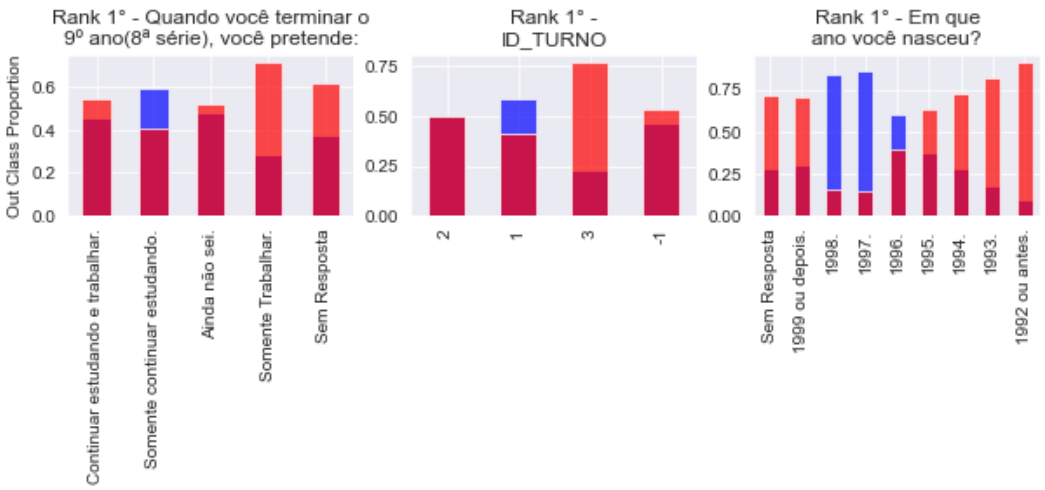
Feature Proportion - In Class - 2



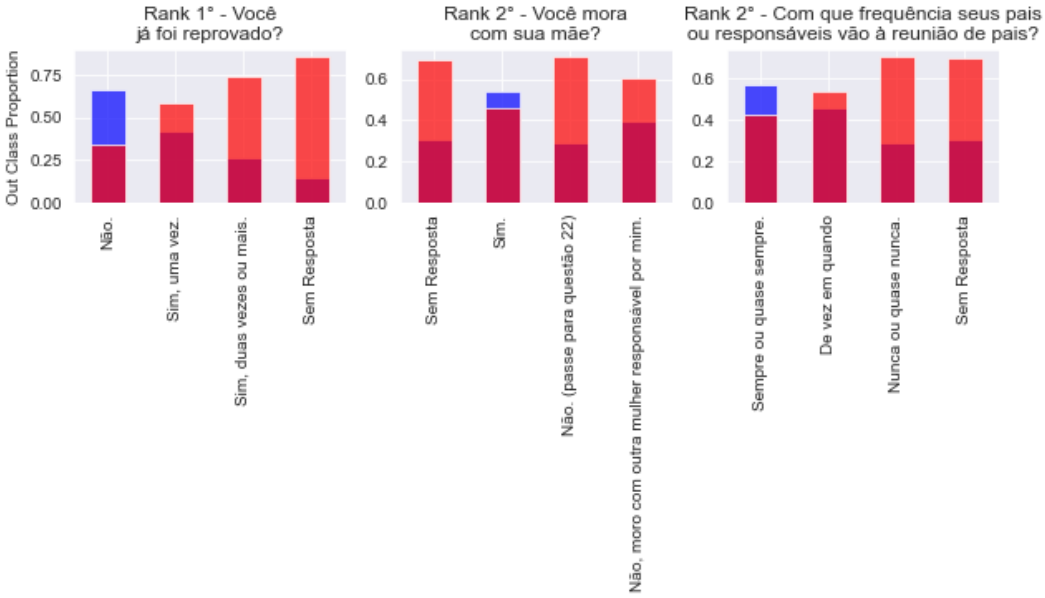
Feature Proportion - In Class - 3



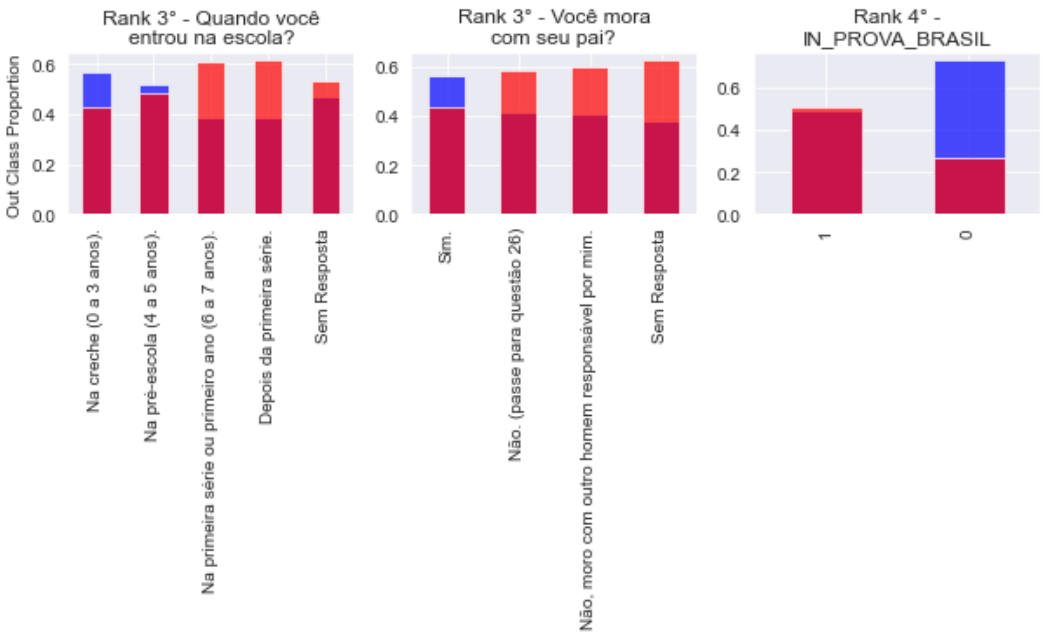
Feature Proportion - In Class - 4



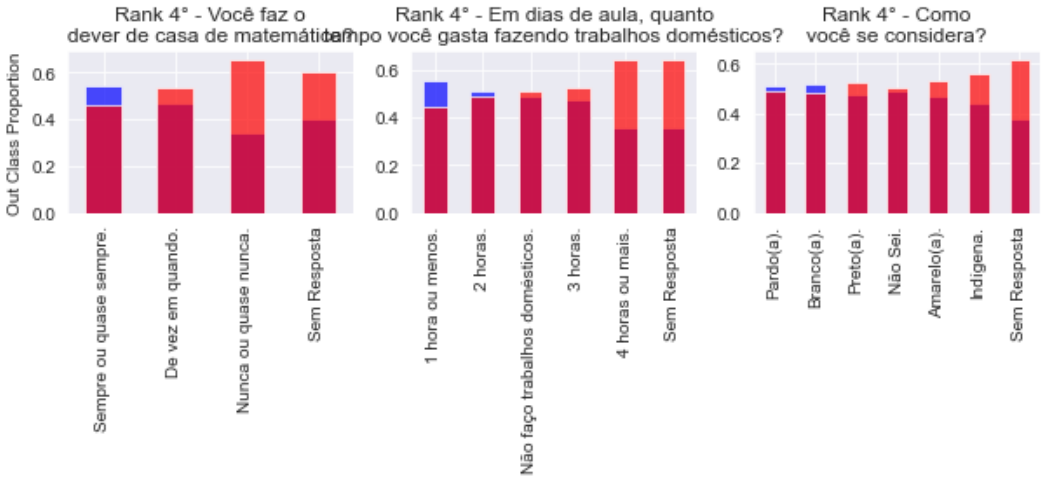
Feature Proportion - Out Class - 1



Feature Proportion - Out Class - 2



Feature Proportion - Out Class - 3



Feature Proportion - Out Class - 4

Feature Importance

Rank	Count in Top 30	Features				
1°	16	TX_RESP_Q058	ID_TURNO	TX_RESP_Q004	TX_RESP_Q049	
2°	15	TX_RESP_Q018	TX_RESP_Q026			
3°	14	TX_RESP_Q047	DESVIO_PADRAO_MT	PROFICIENCIA_LP	TX_RESP_Q022	DESVIO_PADRAO_LP
4°	13	IN_PROVA_BRASIL	TX_RESP_Q055	TX_RESP_Q045	PESO	TX_RESP_Q002
5°	12	TX_RESP_Q030	TX_RESP_Q008	TX_RESP_Q048	TX_RESP_Q052	
6°	11	TX_RESP_Q028	TX_RESP_Q031	TX_RESP_Q034	TX_RESP_Q036	TX_RESP_Q006
7°	10	TX_RESP_Q053	TX_RESP_Q039	TX_RESP_Q056		
8°	8	TX_RESP_Q024	TX_RESP_Q027	ID_ESCOLA		
9°	7	ID_MUNICIPIO				
10°	6	TX_RESP_Q019	ID_TURMA	TX_RESP_Q003		
11°	5	PROFICIENCIA_MT	TX_RESP_Q012	TX_RESP_Q046		
12°	4	ID_UF	TX_RESP_Q017			

Feature Importance - Feature Rank Head

Rank	questão	Enunciado	A	B	C	D	E	F	G	H
1°	TX_RESP_Q058	Quando você terminar o 9º ano(8ª série), você pretende:	Somente continuar estudando.	Somente Trabalhar.	Continuar estudando e trabalhar.	Ainda não sei.				
1°	TX_RESP_Q004	Em que ano você nasceu?	1999 ou depois.	1998.	1997.	1996.	1995.	1994.	1993.	1992 ou antes.
1°	TX_RESP_Q049	Você já foi reprovado?	Não.	Sim, uma vez.	Sim, duas vezes ou mais.					
2°	TX_RESP_Q018	Você mora com sua mãe?	Sim.	Não. (passe para questão 22)	Não, moro com outra mulher responsável por mim.					
2°	TX_RESP_Q026	Com que frequência seus pais ou responsáveis vão à reunião de pais?	Sempre ou quase sempre.	De vez em quando.	Nunca ou quase nunca.					
3°	TX_RESP_Q047	Quando você entrou na escola?	Na creche (0 a 3 anos).	Na pré-escola (4 a 5 anos).	Na primeira série ou primeiro ano (6 a 7 anos).	Depois da primeira série.				
3°	TX_RESP_Q022	Você mora com seu pai?	Sim.	Não. (passe para questão 26)	Não, moro com outro homem responsável por mim.					
4°	TX_RESP_Q055	Você faz o dever de casa de matemática?	Sempre ou quase sempre.	De vez em quando.	Nunca ou quase nunca.					
4°	TX_RESP_Q045	Em dias de aula, quanto tempo você gasta fazendo trabalhos domésticos?	1 hora ou menos.	2 horas.	3 horas.	4 horas ou mais.	Não faço trabalhos domésticos.			
4°	TX_RESP_Q002	Como você se considera?	Branco(a).	Pardo(a).	Preto(a).	Amarelo(a).	Indígena.	Não Sei.		

Feature Importance - Selected Questions