


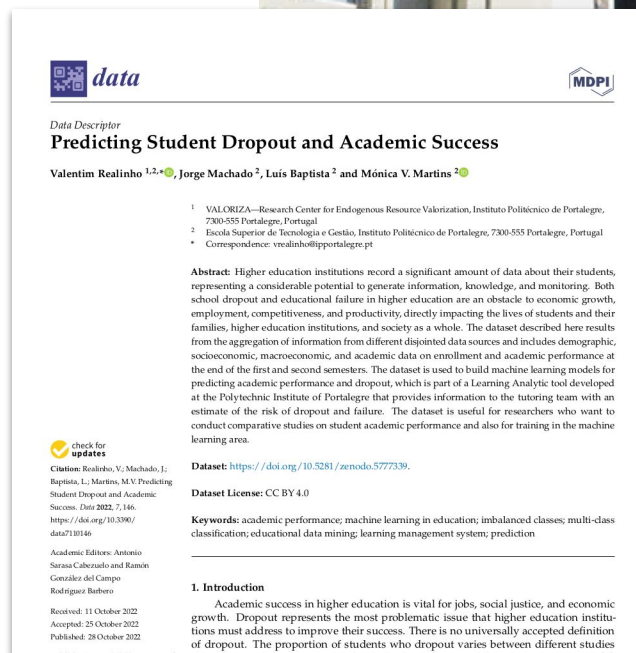
Predição de evasão acadêmica no Instituto Politécnico de Portalegre (IPP)

Leonardo Azzi Martins¹, Matheus Henrique Sabadin¹

¹ Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil
lamartins@inf.ufrgs.br, matheushs15@hotmail.com

1.1 Problema e coleta de dados

- A **evasão acadêmica** pode ser influenciada por uma variedade de fatores, como características demográficas, socioeconômicas, desempenho acadêmico, saúde, entre outras
- O Instituto Politécnico de Portalegre  criou um dataset, a partir de bases disjuntas de dados, sobre estudantes de diversos cursos de graduação [1]



2.1 Análise Exploratória de Dados

Organizou-se a exploração em torno das seguintes perguntas norteadoras:

- **P1. Qual a quantidade e tipos de atributos? Existem inconsistências?**
- **P2. Qual a distribuição do atributo alvo?**
- **P3. Quais os padrões e anomalias dos atributos individuais?**
- **P4. Quais os padrões e anomalias entre os atributos?**

2.1 Análise Exploratória de Dados

P1. Qual a quantidade e tipos de atributos? Existem inconsistências?

- 4424 registros, período 2010-2020;
- 35 atributos: demográficos, socioeconômicos, acadêmicos e macroeconômicos;
- Alguns atributos categóricos estão codificados como *int64* no conjunto de dados; decodificamos estes atributos para sua categoria em *string*
- Resumo dos tipos de atributos
 - 8 categóricos/nominais
 - 10 categóricos/binários
 - 15 numéricos/contínuos
 - 2 numéricos/ordinais (discretos).

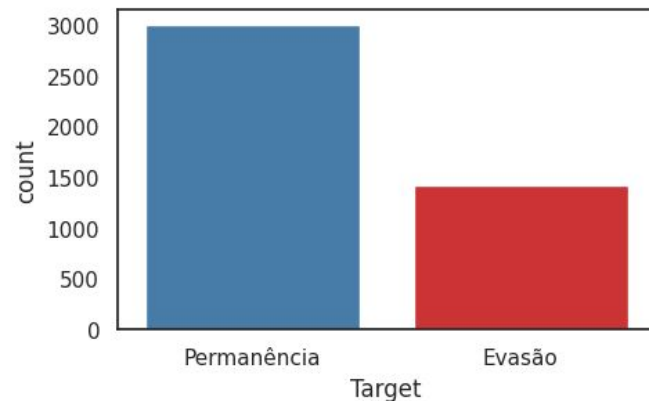
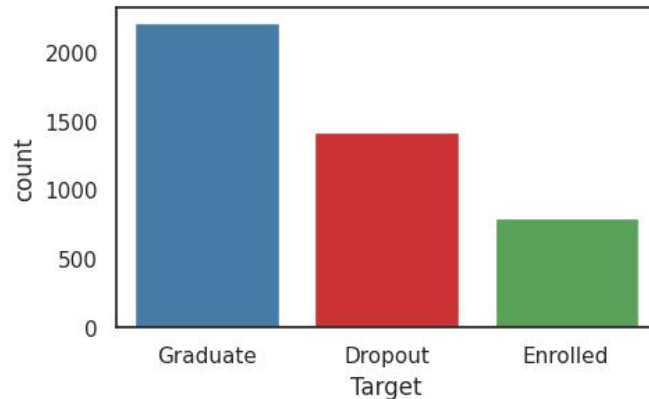
ANEXO 1 - DICIONÁRIO DE DADOS COMPLETO

Dados demográficos			
Atributo	Domínio	Descrição	Categoria(s)
Gender	Categórico binário (codificado)	Gênero binário	0: Feminino 1: Masculino
Marital Status	Categórico nominal (codificado)	Indica o estado civil do estudante	1: Single 2: Married 3: Widower 4: Divorced 5: Facto union 6: Legally separated
Nacionalidade	Categórico nominal (codificado)	Indica a nacionalidade do estudante	1: Portuguese 2: German 3: Spanish 4: Italian 5: Dutch 6: English 7: Lithuanian 8: Angolan 9: Cape Verdean 10: Guinean 11: Mozambican 12: Santomean 13: Turkish 14: Brazilian 15: Romanian 16: Moldova (Republic of) 17: Mexican 18: Ukrainian 19: Russian 20: Cuban 21: Colombian
Age of enrollment	Numérico discreto	Idade que entrou na universidade	N
Dados socioeconômicos			
Atributo	Domínio	Descrição	Categoria(s)
Mother's qualification / Father's	Categórico nominal (codificado)	O grau de formação dos pais do estudante	1—Secondary Education—12th Year of Schooling or Equivalent 2—Higher Education—bachelor's degree

2.1 Análise Exploratória de Dados

P2. Qual a distribuição do atributo alvo?

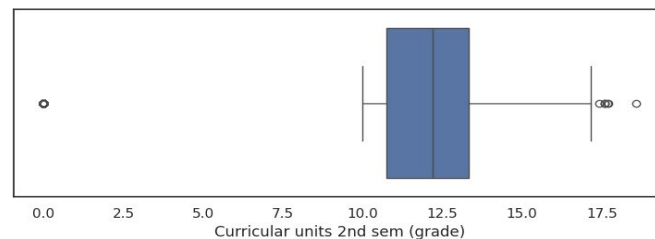
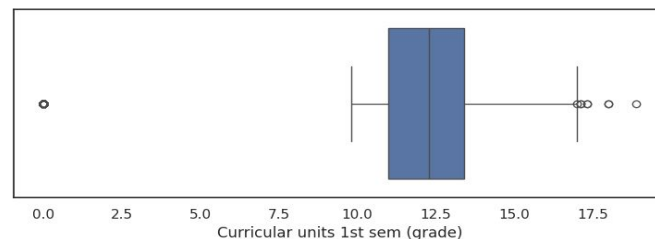
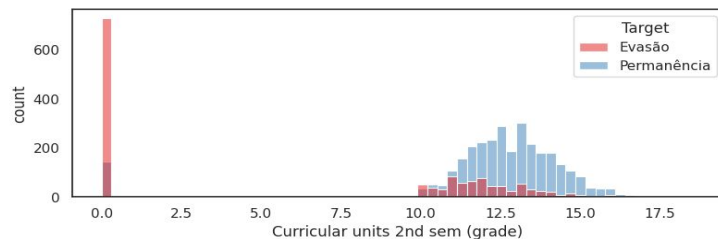
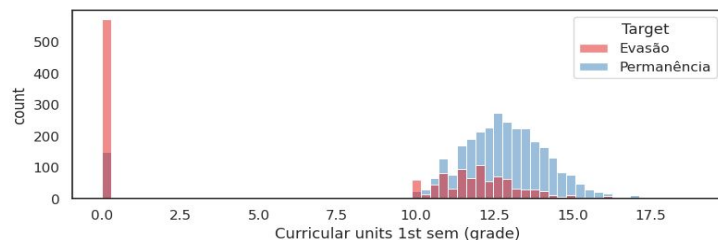
- Atributo alvo: *Target*
- Classes: *Graduate*, *Enrolled*, *Dropout*
- Tendo em vista o objetivo da predição de evasão, as classes *Graduate* e *Enrolled* foram integradas na classe **Permanência**, e a classe *Dropout* foi remapeada para **Evasão**
- Desbalanceamento dos dados: mais alunos em Permanência (quase 68%).



2.1 Análise Exploratória de Dados

P3. Quais os padrões e anomalias dos atributos individuais?

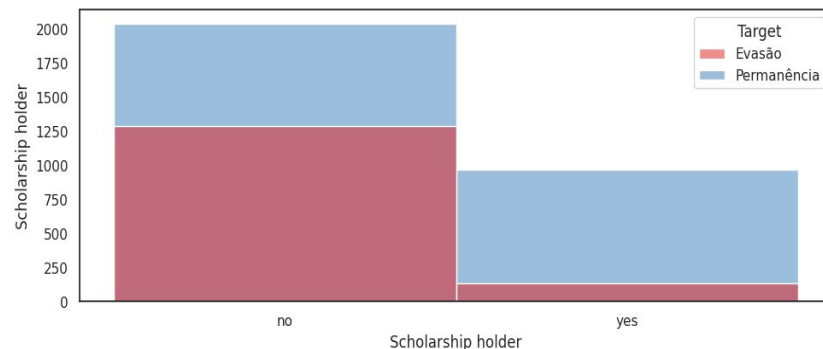
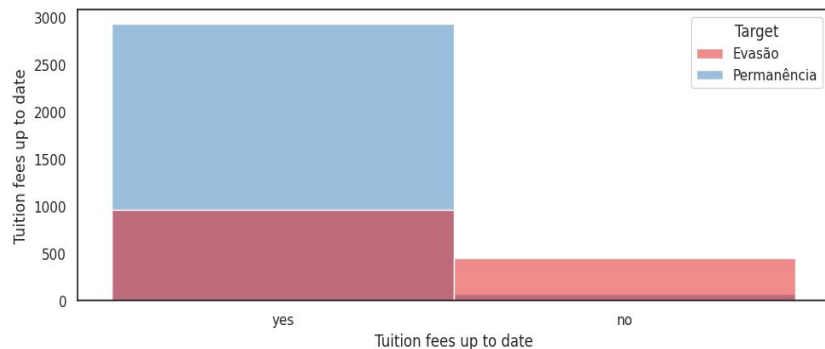
- Dos atributos numéricos, as **notas dos dois primeiros semestres** se destacam para separação dos dados, ainda que não seja uma separação homogênea.



2.1 Análise Exploratória de Dados

P3. Quais os padrões e anomalias dos atributos individuais?

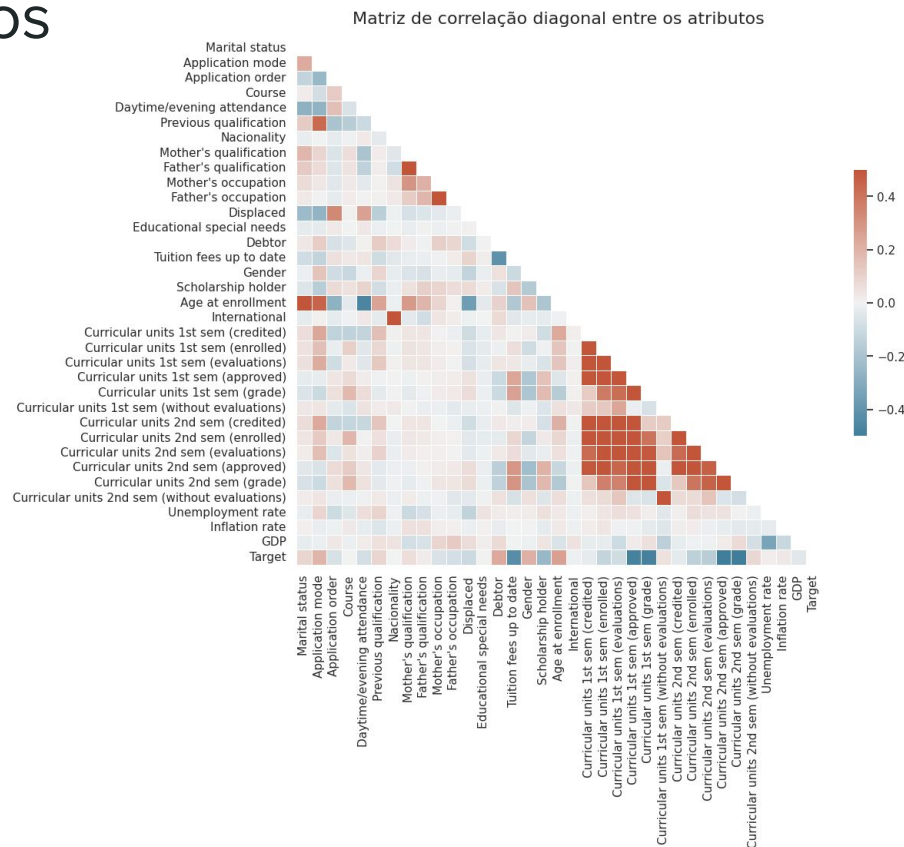
- Dos atributos categóricos, o pagamento em dia da mensalidade e o aluno ser bolsista se destacam para a separação dos dados;
- Em alguns destes atributos existem classes com poucas instâncias



2.1 Análise Exploratória de Dados

P4. Quais os padrões e anomalias entre os atributos?

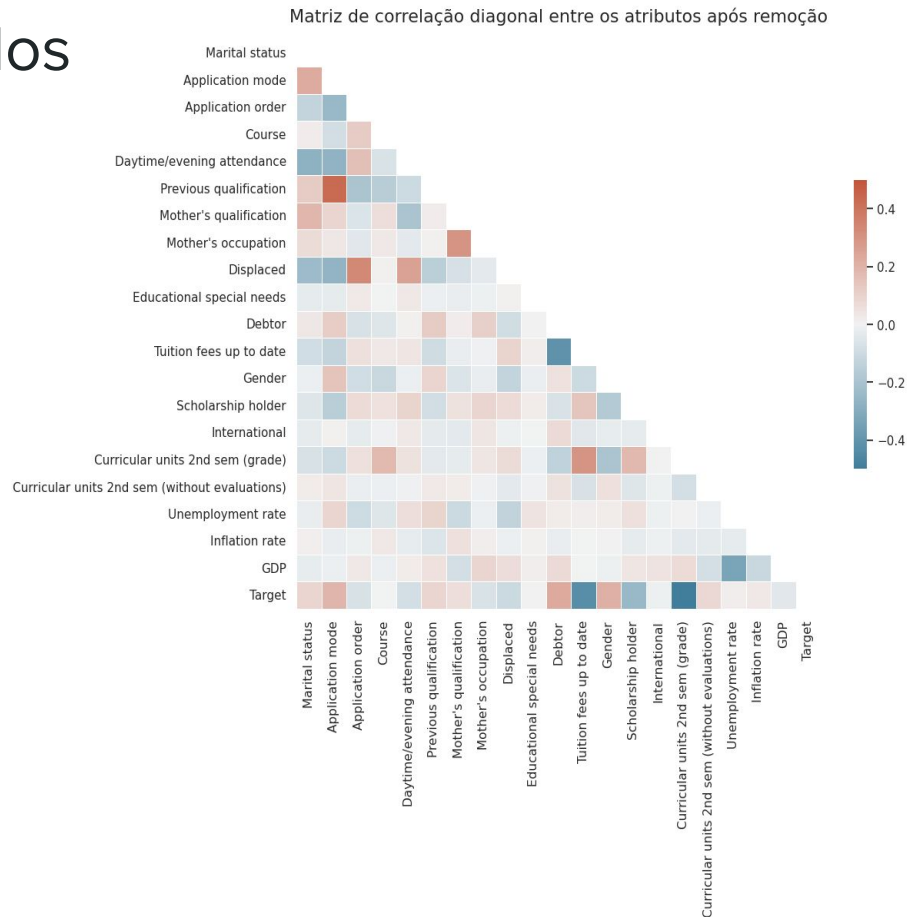
- Utilizada correlação de Pearson;
- Grande parte das correlações com o atributo alvo são fracas ou moderadas;
- Existe correlação moderada entre pares de atributos, indicando que podem ser redundantes;
- Pares redundantes:
 - Age at enrollment / Marital status
 - Age at enrollment / Application mode
 - Mother's occupation / Father's occupation
 - Mother's qualification / Father's qualification
 - International / Nacionalidade
 - Curricular units 1st sem / Curricular units 2nd sem



2.1 Análise Exploratória de Dados

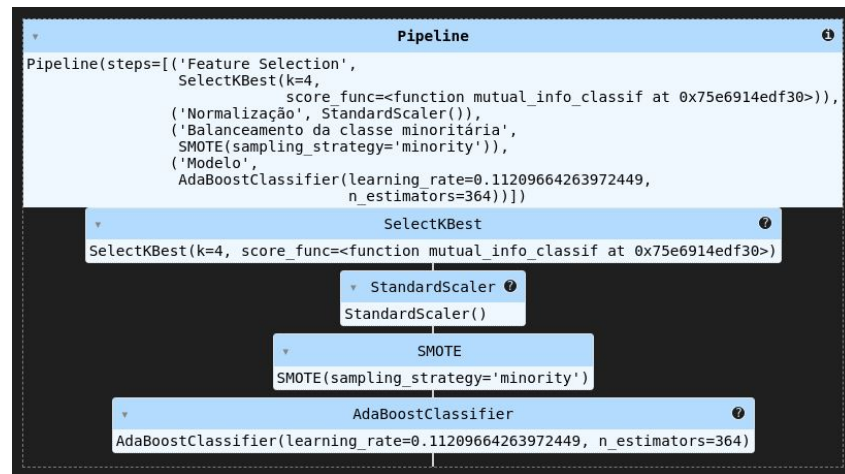
P4. Quais os padrões e anomalias entre os atributos?

- Dos pares redundantes, foram removidos um atributo para cada par:
 - 'Age at enrollment'
 - 'Father's occupation'
 - 'Father's qualification'
 - "Nationality"
 - 'Curricular units 1st sem (grade)', 'Curricular units 1st sem (without evaluations)'
 - Todas relacionadas a 'enrolled', 'evaluations', 'credited' e 'approved' de ambos os semestres.



2.2 Pré-processamento de dados

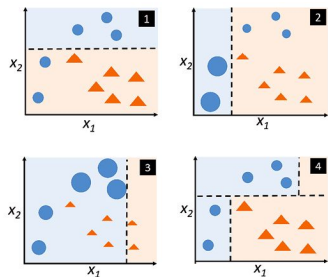
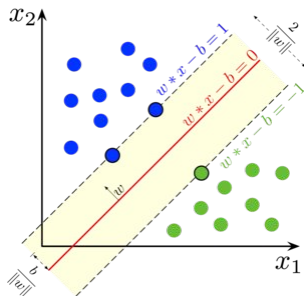
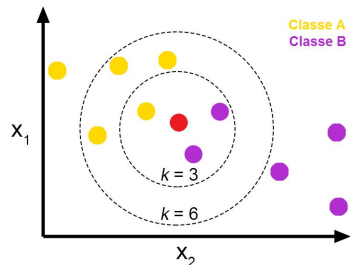
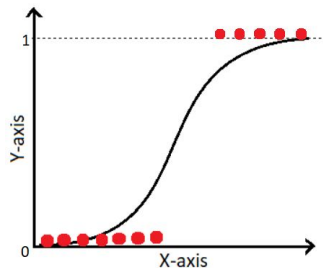
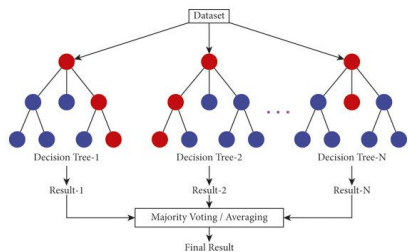
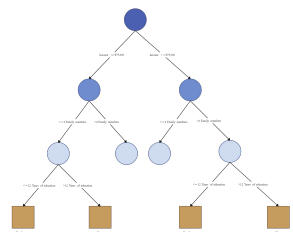
- Criou-se uma *pipeline*, para que o pré-processamento possa ser feito apenas com base nos dados de treinamento, evitando 'vazamento' de dados
- Seleção de atributos
 - *SelectKBest* (k=4), com função de pontuação *mutual_info_classif*;
 - *Application mode, Tuition fees up to date, Scholarship holder, Curricular units 2nd sem (grade).*
- Normalização
 - *StandardScaler*;
- Balanceamento
 - SMOTE, estratégia minoritária



3.1 Seleção de algoritmos

9 algoritmos testados no *Spot-Checking*:

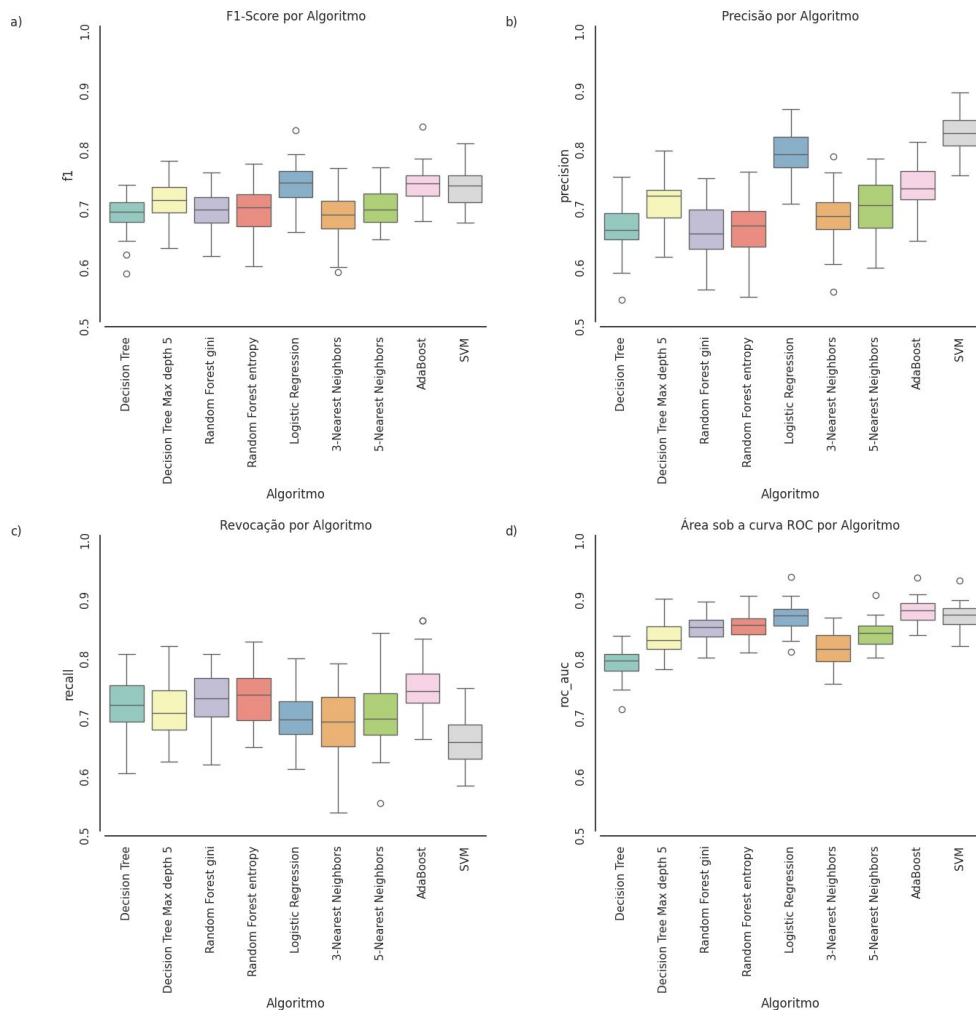
- Árvore de Decisão;
- Árvore de Decisão com profundidade máxima 5;
- Floresta Aleatória (Gini);
- Floresta Aleatória (Entropia);
- Regressão Logística;
- 3-NN;
- 5-NN;
- AdaBoost (kernel linear);
- SVM (50 estimadores, taxa de aprendizado = 1, SAMME)



4. Spot-checking

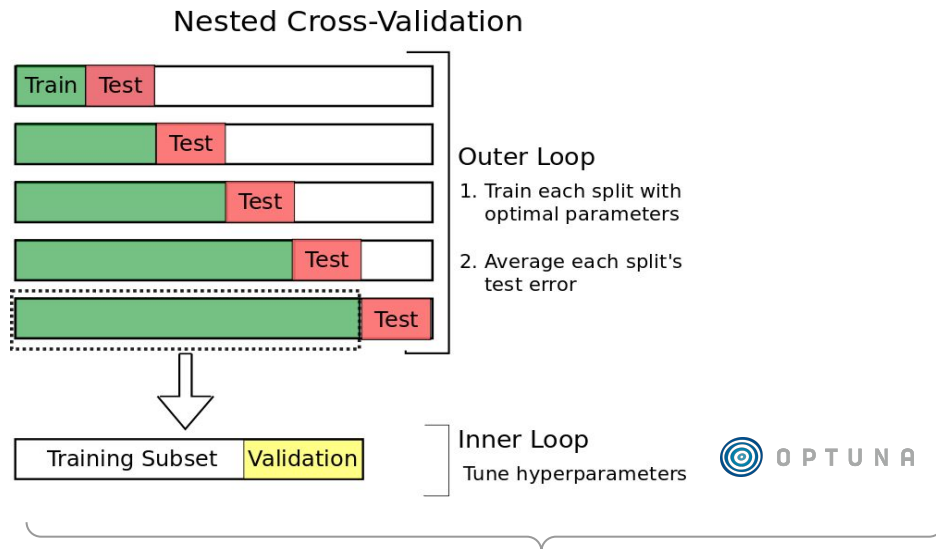
- **Métricas utilizadas:**
 - **F1-Score (principal)**
 - Precisão
 - Revocação
 - ROC AUC
- Estratégia de avaliação:
 - Validação cruzada 10-fold, repetida com 5 sementes aleatórias distintas
- Seleção de 3 estratégias com melhor métrica principal:

- **Regressão Logística**
- **AdaBoost**
- **SVM**



5. Otimização de hiperparâmetros

- Partição inicial de 70-30, onde 30% foi reservado para teste posterior com os hiperparâmetros selecionados
- Framework utilizado: **Optuna**
- Validação cruzada aninhada (Nested CV) para avaliar o desempenho para os hiperparâmetros otimizados
 - 10 folds externos, 50 *trials* e 5 folds internos.



Algoritmo	Hiperparâmetro	Tipo	Espaço de busca
Regressão Logística	C	float (log)	[10-3, 103]
	solver_penalty	categorical	[liblinear_l1, liblinear_l2, lbfgs_l2]
AdaBoost	n_estimators	int	[50, 500]
	learning_rate	float (log)	[0.01, 2.00]
SVM	C	float (log)	[10-4, 102]
	kernel	categorical	[linear, rbf, sigmoid]

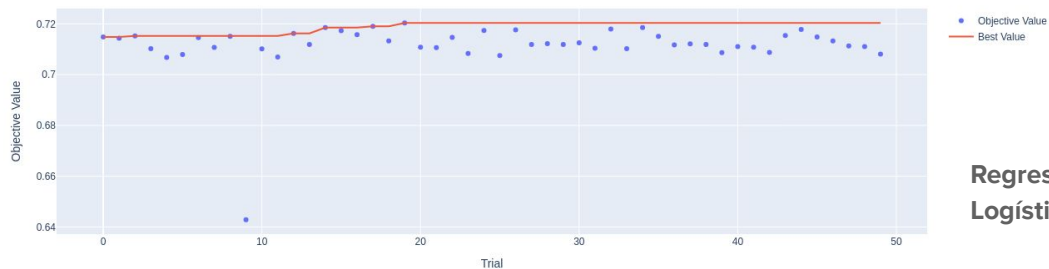
5. Otimização de hiperparâmetros

- Atingiu o melhor valor muito rapidamente;
- A variação dos hiperparâmetros sempre ficava próxima do ótimo;

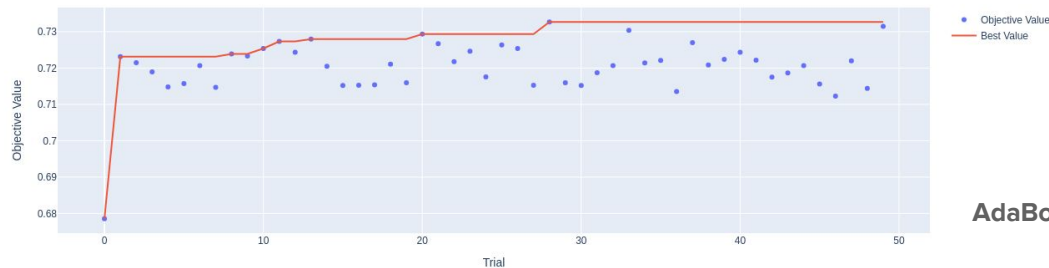
Hiperparâmetros testados:

- **Regressão Logística:**
 - C, solver penalty;
- **AdaBoost:**
 - n_estimators, learning_rate;
- **SVM:**
 - C, kernel (linear, rbf, sigmoid).

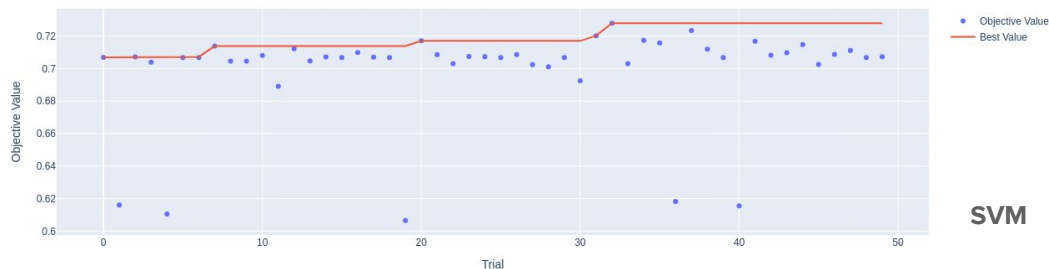
Optimization History Plot



Regressão Logística



AdaBoost



SVM

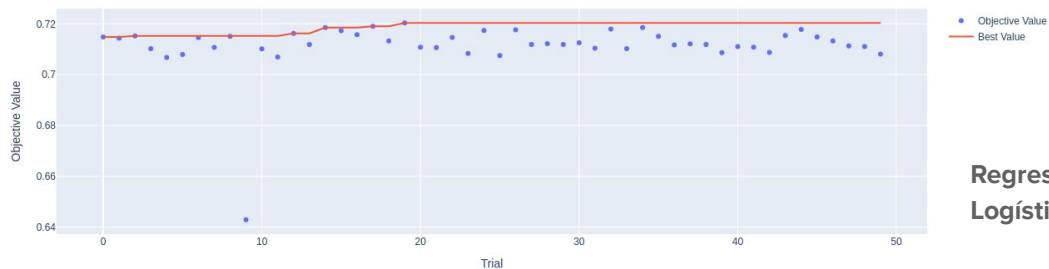
5. Otimização de hiperparâmetros

- Atingiu o melhor valor muito rapidamente;
- A variação dos hiperparâmetros sempre ficava próxima do ótimo;

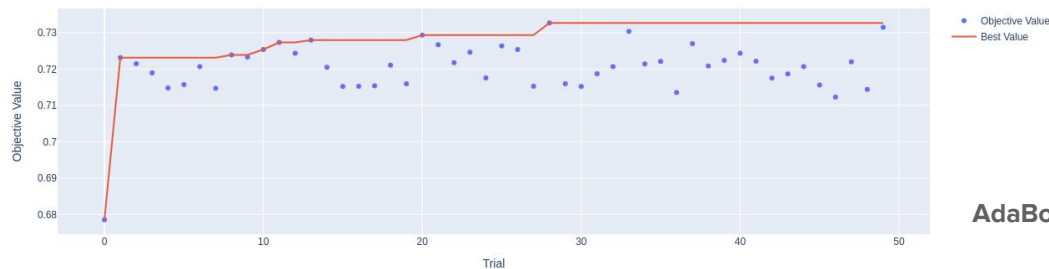
Hiperparâmetros testados:

- **Regressão Logística:**
 - C, solver penalty;
- **AdaBoost:**
 - n_estimators, learning_rate;
- **SVM:**
 - C, kernel (linear, rbf, sigmoid).

Optimization History Plot



Regressão Logística



AdaBoost

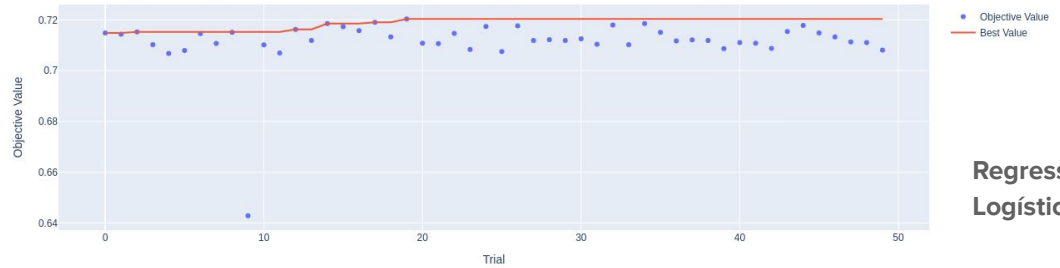


Em poucas tentativas, a otimização encontra um valor ótimo, mantendo poucas variações ao longo das tentativas. Isto pode indicar uma simplicidade do próprio conjunto de dados.

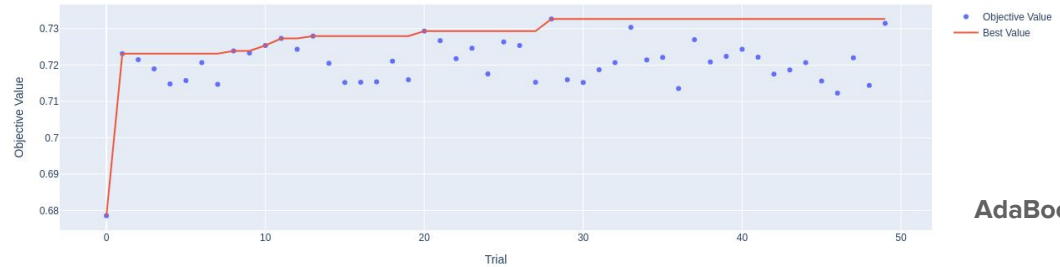


**Faz alguma
coisa aí bicho...**

Optimization History Plot



**Regressão
Logística**



AdaBoost

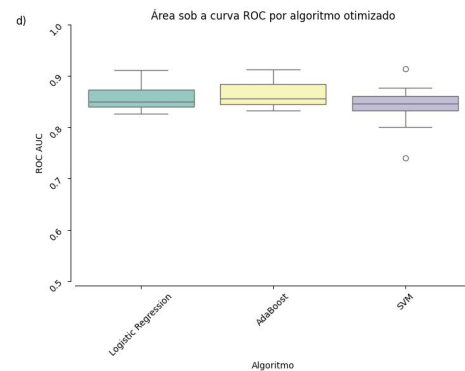
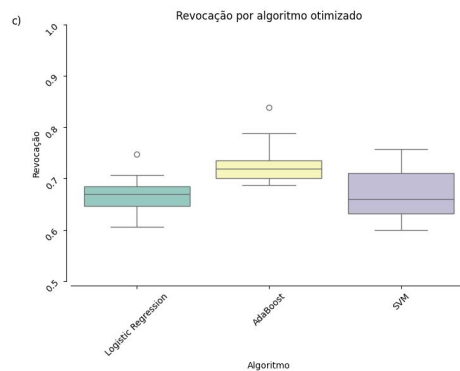
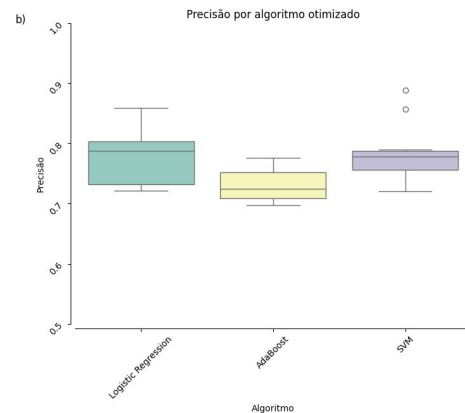
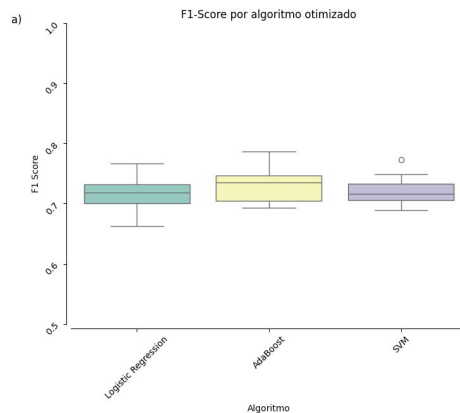


Em poucas tentativas, a otimização encontra um valor ótimo, mantendo poucas variações ao longo das tentativas. Isto pode indicar uma simplicidade do próprio conjunto de dados.

5.1 Desempenho

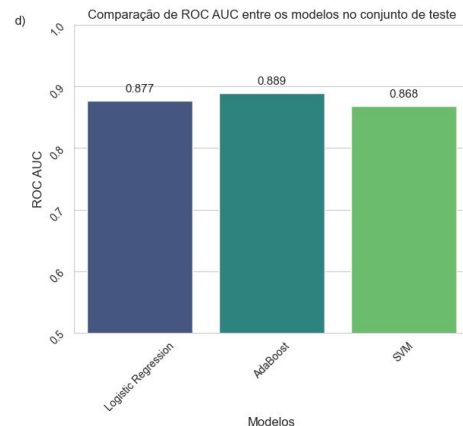
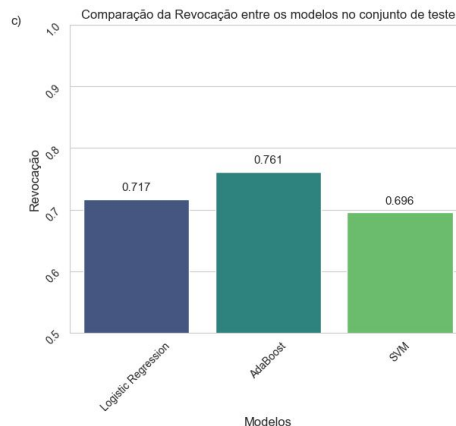
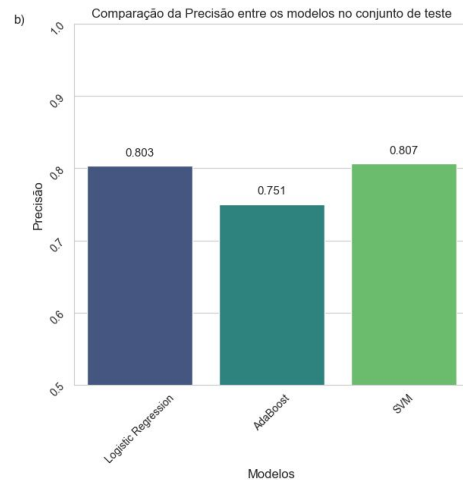
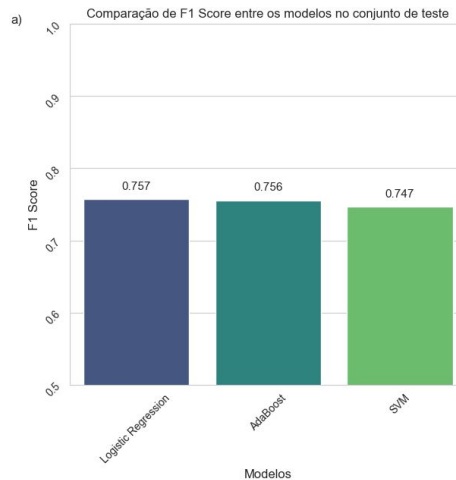
- A otimização de hiperparâmetros, em geral, apresentou baixa variância na distribuição de desempenho
- Foram selecionados os hiperparâmetros com a melhor pontuação na métrica principal.

Algoritmo	F1-Score	Hiperparâmetro	Valor otimizado
Regressão Logística	0.766839	C	0.0661402819042377
		solver_penalty	default
AdaBoost	0.786730	n_estimators	258
		learning_rate	0.5686136448966946
Máquina de Vetores de Suporte	0.773196	C	18.72430564502587
		kernel	rbf



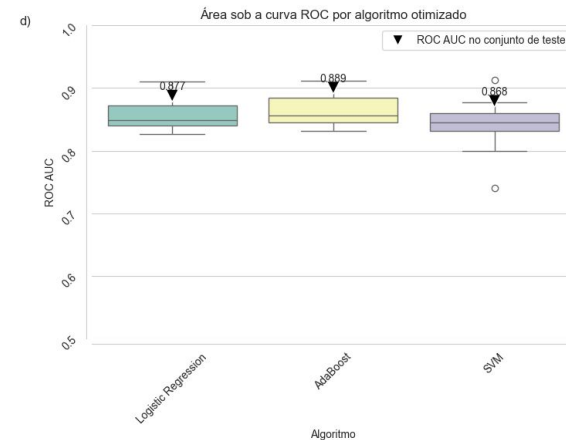
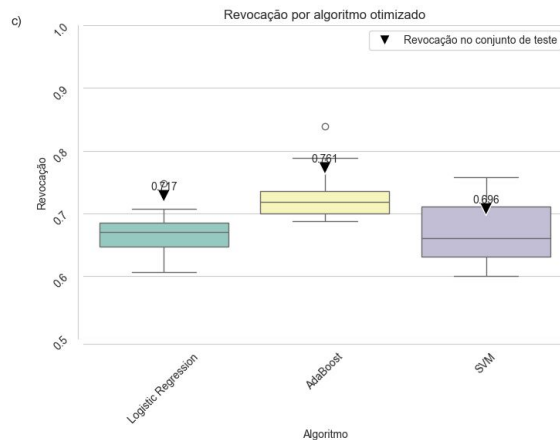
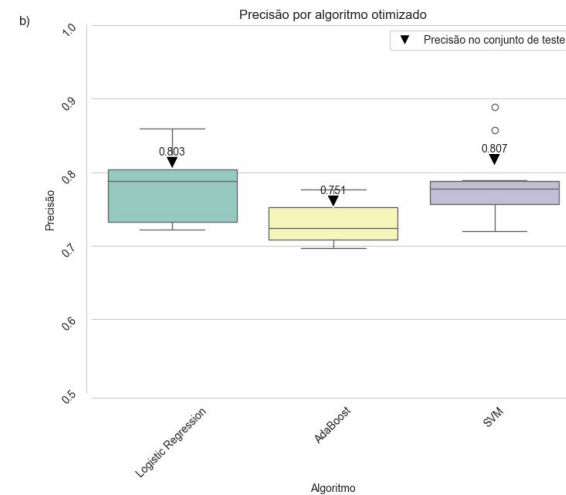
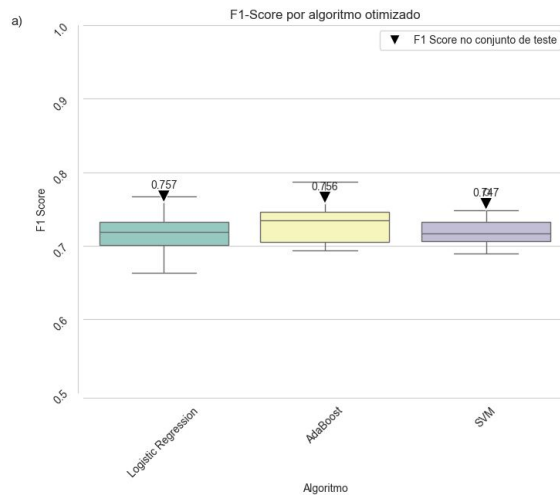
6. Comparação

- Os algoritmos com seus melhores hiperparâmetros foram treinados na partição inicial de treino e validados na partição reservada de teste;
- Os desempenhos, em geral, são similares entre Regressão Logística, AdaBoost e SVM;
- F1-Score próximo de 0.75;
- Tendências semelhantes ao que foi visto no *Spot-Checking*.

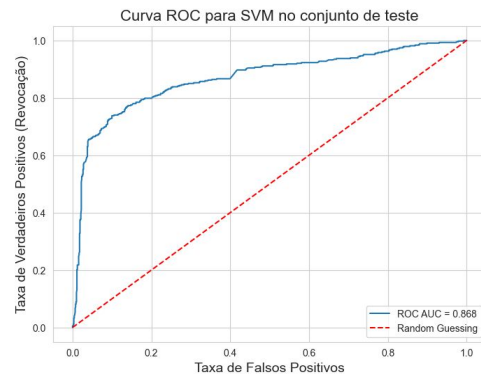
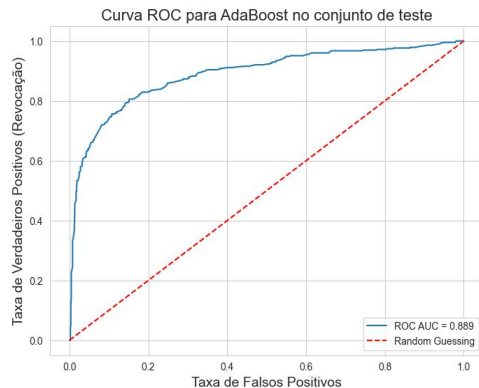
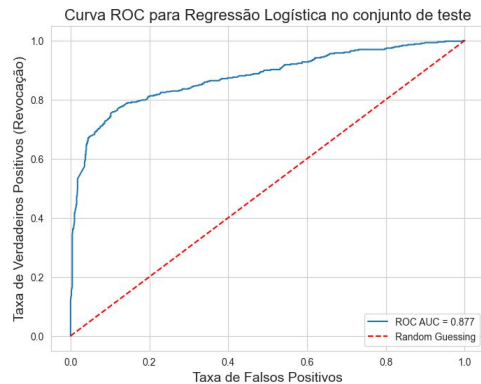


6. Comparação

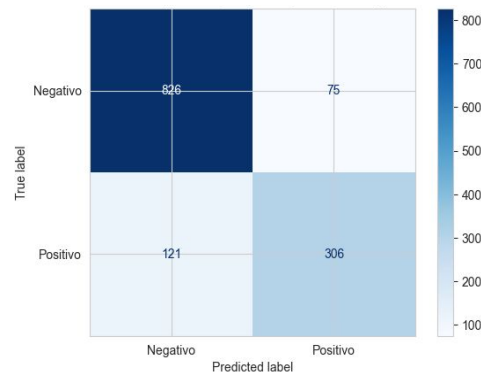
- **AdaBoost** foi selecionado como algoritmo final, desempatado devido ao seu desempenho superior na métrica ROC AUC.
- As pontuações no conjunto de teste (▼) são comparadas com a distribuição de desempenho na otimização
 - Diferenças entre os modelos são pequenas. (Eixo Y entre 0.5 e 1.0).



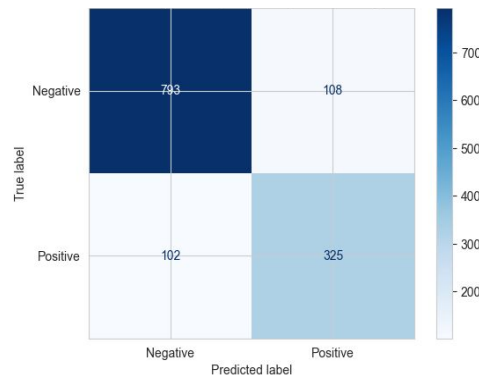
6. Comparação



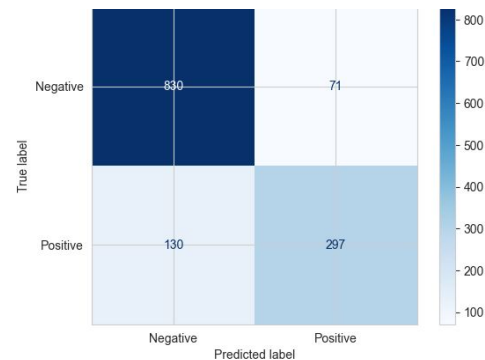
Matriz de confusão para Regressão Logística no conjunto de teste



Matriz de confusão para AdaBoost no conjunto de teste



Matriz de confusão para SVM no conjunto de teste

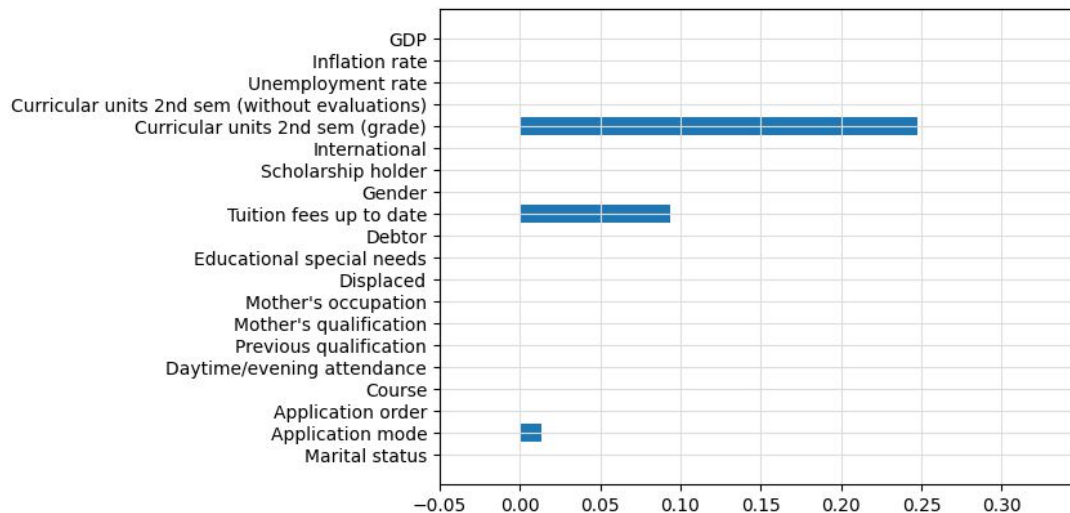


7.1 Interpretação: importância de atributos

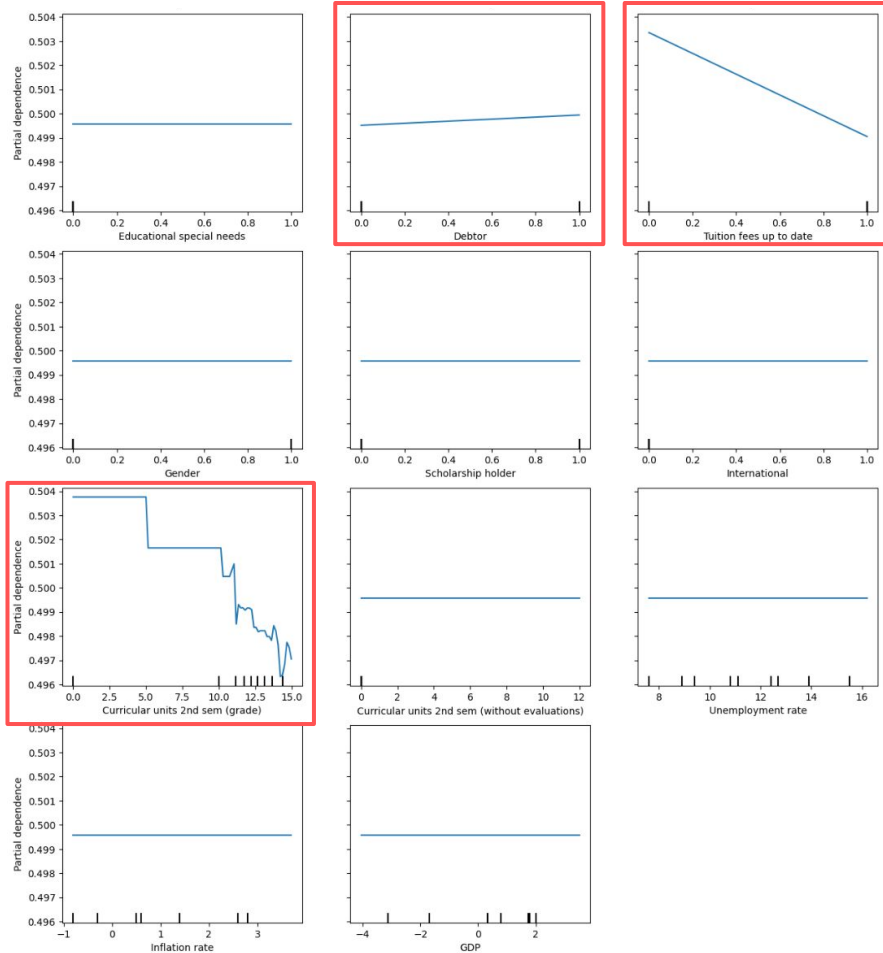
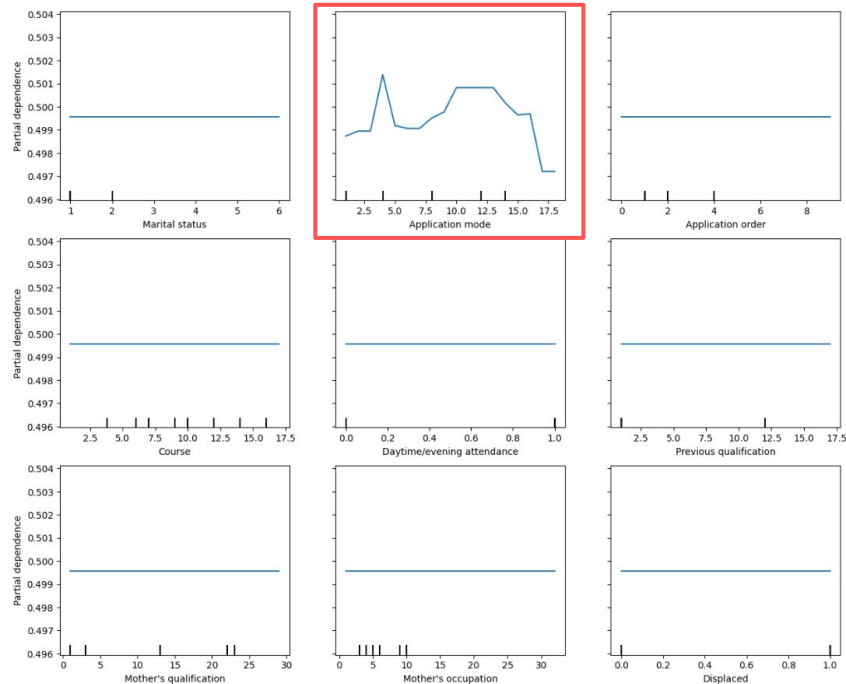
Atributos mais importantes:

- Notas do 2º semestre;
- Pagamento das "taxas acadêmicas" em dia;
- Modo de aplicação.

Reforça a análise feita no EDA de que poucos atributos poderiam contribuir significativamente.

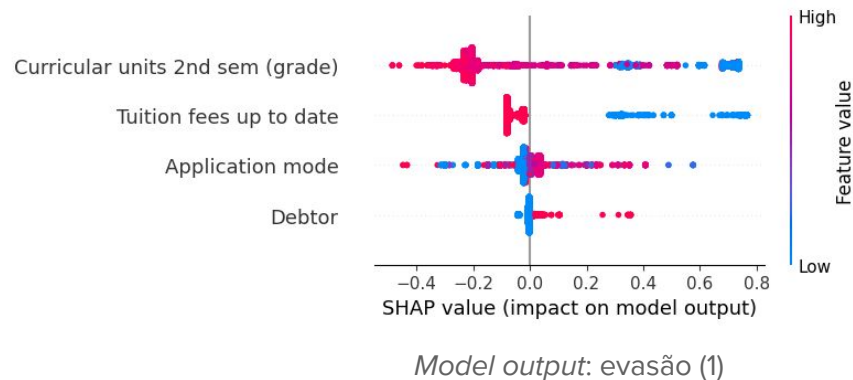


7.2 Interpretação: PDP



7.3 Interpretação: SHapley Additive exPlanations (SHAP)

- Curricular units 2nd sem (grade):
 - Notas altas → Contribuição negativa ou positiva $[-0.5, 0.6]$
 - Notas baixas → Contribuição positiva alta $[0.3, 0.8]$
- Tuition fees up to date:
 - Em dia (1) → Contribuição negativa pequena $[-0.1, 0.0]$
 - Atrasado (0) → Contribuição positiva alta $[0.2, 0.8]$
- Application mode:
 - Existem muitas intersecções entre valores, pois codificam outras categorias binárias, dificultando sua explicação
- Debtor:
 - Devedor → Contribuição esparsa positiva $[0.0, 0.4]$
 - Não devedor → Contribuição concentrada em torno de 0.0
 - Impactos menores, mas bem delimitados para a separação entre evasão e permanência



8. Considerações finais

- Os poucos atributos selecionados já oferecem boa separação entre evasão e permanência;
- Modelo final escolhido é o AdaBoost:
 - Desempenho: $F1 = 0.75$, $ROC\ AUC = 0.89$;
- Análise mostra que **atributos financeiros e acadêmicos** são determinantes na evasão.
- Possibilidades Futuras:
 - Explorar novos algoritmos como *XGBoost* ou Redes Neurais.
 - Coletar novos dados relacionados à saúde e desempenho em outros semestres.
 - Decodificar o atributo categórico *Application Mode* com técnicas de transformação como *one-hot encoding*.

Referências

- [1] REALINHO, V.; MACHADO, J.; BAPTISTA, L.; MARTINS, M. V. Predicting Student Dropout and Academic Success. *Data*, v. 7, n. 146, 2022. DOI: <https://doi.org/10.3390/data7110146>.
- [2] Predict students' dropout and academic success. Disponível em: <https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention/data>. Acesso em: 14 dez. 2024.
- [3] SCHOBER, P.; BOER, C.; SCHWARTE, L. A. Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, v. 126, n. 5, p. 1763-1768, maio 2018. DOI: 10.1213/ANE.0000000000002864. Disponível em: https://journals.lww.com/anesthesia-analgesia/fulltext/2018/05000/correlation_coefficients_appropriate_use_and_50.aspx. Acesso em: 14 dez. 2024.
- [4] Machine Learning Modeling Pipelines. Disponível em: <https://machinelearningmastery.com/machine-learning-modeling-pipelines/>. Acesso em: 14 dez. 2024.
- [5] SKLEARN. Feature Selection - SelectKBest. Disponível em: https://scikit-learn.org/1.6/modules/generated/sklearn.feature_selection.SelectKBest.html. Acesso em: 14 dez. 2024.
- [6] SKLEARN. Feature Selection - Mutual Info Classif. Disponível em: https://scikit-learn.org/1.6/modules/generated/sklearn.feature_selection.mutual_info_classif.html#sklearn.feature_selection.mutual_info_classif. Acesso em: 14 dez. 2024.
- [7] Optimizing Performance: SelectKBest for Efficient Feature Selection in Machine Learning. Medium. Disponível em: <https://medium.com/@Kavya2099/optimizing-performance-selectkbest-for-efficient-feature-selection-in-machine-learning-3b635905ed48>. Acesso em: 14 dez. 2024.
- [8] SKLEARN. StandardScaler. Disponível em: <https://scikit-learn.org/1.6/modules/generated/sklearn.preprocessing.StandardScaler.html>. Acesso em: 14 dez. 2024.
- [9] IMBALANCED-LEARN. SMOTE. Disponível em: https://imbalanced-learn.org/0.12/references/generated/imblearn.over_sampling.SMOTE.html. Acesso em: 14 dez. 2024.
- [10] SKLEARN. Logistic Regression. Disponível em: https://scikit-learn.org/1.6/modules/generated/sklearn.linear_model.LogisticRegression.html. Acesso em: 14 dez. 2024.
- [11] GEEKSFORGEEEKS. Regularization in Machine Learning. Disponível em: <https://www.geeksforgeeks.org/regularization-in-machine-learning/>. Acesso em: 14 dez. 2024.
- [12] SKLEARN. AdaBoostClassifier. Disponível em: <https://scikit-learn.org/1.6/modules/generated/sklearn.ensemble.AdaBoostClassifier.html>. Acesso em: 14 dez. 2024.
- [13] GEEKSFORGEEEKS. Radial Basis Function Kernel in Machine Learning. Disponível em: <https://www.geeksforgeeks.org/radial-basis-function-kernel-machine-learning/#radial-basis-function-kernel>. Acesso em: 14 dez. 2024.
- [14] GEEKSFORGEEEKS. Creating Linear Kernel SVM in Python. Disponível em: <https://www.geeksforgeeks.org/creating-linear-kernel-svm-in-python/>. Acesso em: 14 dez. 2024.
- [15] SKLEARN. SVM Kernel Functions. Disponível em: <https://scikit-learn.org/1.6/modules/svm.html#kernel-functions>. Acesso em: 14 dez. 2024.
- [16] SKLEARN. SVC. Disponível em: <https://scikit-learn.org/1.6/modules/generated/sklearn.svm.SVC.html>. Acesso em: 14 dez. 2024.
- [17] MENDONZA, Mariana R. Interpretação de Modelos Preditivos. Apresentação de Slides, 2024. Acesso em: 14 dez. 2024.

Obrigado!

Perguntas?

