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

Leonardo Azzi Martins[1], Matheus H. Sabadin[1]

[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. Identificação do Problema e Coleta de Dados

1.1 Identificação do Problema

A evasão acadêmica de estudantes universitários pode ser influenciada por uma variedade de fatores, incluindo características demográficas, socioeconômicas e desempenho acadêmico prévio. Uma ferramenta estatística que dê recursos para compreender e prever o sucesso ou a desistência dos estudantes pode auxiliar instituições de ensino a implementar estratégias para melhorar a retenção e o desempenho estudantil.

Este trabalho tem como objetivo desenvolver um modelo baseado em Aprendizado de Máquina para prever a evasão acadêmica dos estudantes, ou seja, se o aluno irá desistir ou permanecer - tanto matriculado quanto formado. A predição antecipada da situação acadêmica permite que as instituições de ensino superior adotem medidas proativas para apoiar os estudantes em risco e aprimorar as políticas educacionais.

1.2 Coleta de Dados

Utilizou-se o conjunto de dados "*Predicting Student Dropout and Academic Success*" [1], publicado em 2022. Este *dataset* abrange um amplo espectro de informações sobre alunos matriculados em diversos cursos de graduação do Instituto Politécnico de Portalegre, uma instituição de ensino superior em Portugal. Ele é composto por dados demográficos dos estudantes, fatores socioeconômicos e variáveis de desempenho acadêmico, coletados no momento da matrícula. Estes incluem o modo de aplicação, estado civil, curso escolhido, entre outros. Fatores macroeconômicos externos como a taxa de desemprego, taxa de inflação e o PIB da região também estão disponíveis, oferecendo dados adicionais que podem ser utilizados para entender contextos mais amplos que afetam os estudantes.

Adicionalmente, o conjunto de dados permite a análise do desempenho estudantil ao término de dois semestres através de métricas sobre unidades curriculares creditadas, inscritas, avaliadas e aprovadas, bem como as notas correspondentes. Além desses dados, o *dataset* contém a situação do aluno no curso, já rotulado como “alvo” preditivo: graduado, desistente ou matriculado. O *dataset* foi disponibilizado no formato csv e adquirido através do *Kaggle* [2]. As informações contidas são úteis para uma análise detalhada das trajetórias acadêmicas dos alunos, facilitando estudos sobre a dinâmica de ingresso, progressão e conclusão dos cursos superiores.

O *dataset* é composto de 4424 registros para 35 atributos. Cada um destes representa um aluno no período entre 2010 e 2020. Com exceção do atributo alvo, todos os demais atributos categóricos já foram codificados. Portanto, dos atributos originais do *dataset*, 10 são numéricos/binários, 5 são numéricos/contínuos, e 20 são numéricos/discretos. Considerando a natureza destes atributos binários e discretos, na verdade 8 são categóricos/nominais, 10 são categóricos/binários, 15 são numéricos/contínuos e 2 são numéricos/ordinais discretos. A seguir, apresentamos o dicionário de dados, descrevendo o significado de cada atributo agrupado por sua natureza.

1. **Dados demográficos:**
    a. *Estado Civil (Marital Status):* Indica o estado civil do estudante;
    b. *Nacionalidade (Nacionality):* Nacionalidade do estudante;

c. *Deslocado (Displaced):* Indica se o estudante é uma pessoa a qual foi obrigada a sair de seu país de residência habitual por motivos de conflito armado, violência, violação de direitos humanos ou desastres naturais;
d. *Gênero (Gender)*: Gênero binário do estudante;
e. *Idade na Inscrição (Age at enrollment)*: Idade do estudante no momento da inscrição;
f. *Estudante Internacional (International)*: Se o estudante é internacional;

2. Dados socioeconômicos

a. *Grau de Ensino da Mãe (Mother's Qualification):* O grau de ensino da mãe do estudante;
b. *Grau de Ensino do Pai (Father's Qualification):* O grau de ensino do pai do estudante;
c. *Ocupação da Mãe (Mother's Occupation):* A ocupação da mãe do estudante;
d. *Ocupação do Pai (Father's Occupation):* A ocupação do pai do estudante;
e. *Necessidades Educacionais Especiais (Educational Special Needs):* Se o estudante possui alguma necessidade educacional especial;
f. *Devedor (Debtor):* Indica se o estudante possui débitos;
g. *Mensalidade em Dia (Tuition fees up to date):* Se as mensalidades do estudante estão atualizadas;
h. *Bolsista (Scholarship holder):* Indica se o estudante é beneficiário de alguma bolsa de estudos;

3. Fatores macroeconômicos

a. *Taxa de desemprego (Unemployment rate):* Taxa de desemprego do país;
b. *Taxa de inflação (Inflation rate):* Inflação do país;
c. *PIB (GDP):* Produto interno bruto do país;

4. Dados acadêmicos durante a matrícula

a. *Modo de Aplicação (Application Mode):* Método utilizado pelo estudante para aplicar a uma vaga na instituição de ensino.
b. *Ordem de Aplicação (Application Order):* Ordem de prioridade numérica da aplicação do estudante.
c. *Curso (Course):* Curso escolhido pelo estudante.
d. *Rotina de Aulas Diurna/Noturna (Daytime/evening Attendance):* Indica se o estudante frequenta aulas durante o dia ou à noite.
e. *Qualificação Prévia (Previous Qualification):* Qualificações obtidas antes do ingresso no ensino superior.

5. Dados acadêmicos ao final do primeiro semestre

a. *Atividades Curriculares no 1º Semestre Creditadas (Curricular units 1st sem (credited)):* Número de atividades curriculares creditadas pelo estudante no primeiro semestre.

b. *Atividades Curriculares no 1º Semestre Inscritas (Curricular units 1st sem (enrolled)):* Número de atividades curriculares nas quais o estudante se inscreveu no primeiro semestre.
c. *Atividades Curriculares no 1º Semestre Avaliadas (Curricular units 1st sem (evaluations)):* Número de atividades curriculares avaliadas pelo estudante no primeiro semestre.
d. *Atividades Curriculares no 1º Semestre Aprovadas (Curricular units 1st sem (approved)):* Número de atividades curriculares aprovadas pelo estudante no primeiro semestre.
e. *Notas nas atividades Curriculares no 1º Semestre (Curricular units 1st sem (grade)):* Notas nas atividades curriculares realizadas pelo estudante.
f. *Atividades Curriculares no 1º Semestre não avaliadas (Curricular units 1st sem (without evaluations)):* Número de atividades curriculares sem avaliação pelo estudante no primeiro semestre.

6. Dados acadêmicos ao final do segundo semestre

a. *Atividades Curriculares no 2º Semestre Creditadas (Curricular units 2st sem (credited)):* Número de atividades curriculares creditadas pelo estudante no segundo semestre.
b. *Atividades Curriculares no 2º Semestre Inscritas (Curricular units 2st sem (enrolled)):* Número de atividades curriculares nas quais o estudante se inscreveu no segundo semestre.
c. *Atividades Curriculares no 2º Semestre Avaliadas (Curricular units 2st sem (evaluations)):* Número de atividades curriculares avaliadas pelo estudante no segundo semestre.
d. *Atividades Curriculares no 2º Semestre Aprovadas (Curricular units 2st sem (approved)):* Número de atividades curriculares aprovadas pelo estudante no segundo semestre.
e. *Notas nas atividades Curriculares no 2º Semestre (Curricular units 2st sem (grade)):* Notas nas atividades curriculares realizadas pelo estudante.
f. *Atividades Curriculares no 2º Semestre não avaliadas (Curricular units 2st sem (without evaluations)):* Número de atividades curriculares sem avaliação pelo estudante no segundo semestre.

7. Situação do aluno

a. *Atributo alvo (Target):* Qual o status do aluno quando durante a criação do *dataset* [1], podendo ser: graduado (*Graduate*), desistente (*Dropout*) ou matriculado (*Enrolled*).

2. Análise Exploratória e Pré-Processamento dos Dados

Esta etapa se divide em dois objetivos: compreender como os atributos do conjunto de dados se relacionam com a probabilidade de evasão acadêmica de estudantes; selecionar os atributos mais relevantes para o objetivo de aprendizado e tratá-los para serem consumidos corretamente por algoritmos de Aprendizado de Máquina (AM).

2.1 Análise Exploratória dos Dados

Realizamos uma Análise Exploratória de Dados com os seguintes objetivos:

- **Compreensão dos dados**: identificação dos tipos de atributos (numéricos/categóricos), distribuição das variáveis e detecção de valores faltantes ou anômalos.
- **Análise do atributo alvo**: verificação da distribuição das classes ("*Graduate*", "*Dropout*", "*Enrolled*") para entender o balanceamento dos dados.
- **Identificação de padrões e relações**: utilização de matrizes de correlação e gráficos para identificar quais atributos possuem maior influência na situação acadêmica dos estudantes.

Organizamos a exploração em torno de perguntas norteadoras, que serviram como guia para a compreensão dos dados:

- **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?

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

O *dataset* contém 4.424 registros, cada um representando um aluno no período de 2010 a 2020, e é composto por 35 atributos, dos quais apenas o atributo alvo não foi codificado. No *dataset* original os atributos foram classificados como 10 numéricos/binários, 5 numéricos/contínuos e 20 numéricos/discretos.

O atributo alvo está codificado como tipo *object*, indicando que armazena ponteiros para *strings*. Os 34 atributos restantes, potencialmente treináveis, estão codificados como valores numéricos *float64* ou *int64*. No entanto, de acordo com a descrição do conjunto de dados, alguns desses atributos são categóricos que foram codificados como *int64*. Isso dificulta a interpretação de gráficos e a análise conjunta de atributos categóricos e numéricos. Para resolver esse problema, reconstruímos os atributos categóricos em um novo *data frame*, com base nas informações fornecidas pelo descritor de dados do *dataset* [1]. Uma descrição detalhada do tipo de cada atributo pode ser encontrada no Anexo 1. Portanto, essa categorização de atributos foi decodificada e remapeada para 8 categóricos/nominais, 10 categóricos/binários, 15 numéricos/contínuos e 2 numéricos/ordinais (discretos).

P2. Qual a distribuição do atributo alvo?

O atributo alvo apresenta um significativo desbalanceamento entre as três classes: são 2.209 instâncias para *'Graduate'*, 794 para *'Enrolled'* e 1.421 para *'Dropout'*. Caso as classes *'Graduate'* e

'Enrolled' sejam agrupadas, o desbalanceamento se torna ainda mais evidente. Nesse cenário, o agrupamento de teria 52,68% do número de instâncias de *'Graduate'* sozinho, com uma diferença absoluta de 1.582 instâncias.

Para uma análise focada na evasão, a separação original entre as classes *'Graduate'* e *'Enrolled'* é pouco relevante, já que em ambos os casos os alunos não desistiram do curso. Além disso, os atributos apresentam correlações fracas ou irrelevantes em relação a essas classes. Como o objetivo do aprendizado é classificar os dados de forma binária entre evasão e permanência, esses rótulos foram unificados: *'Graduate'* e *'Enrolled'* foram agrupados sob o rótulo 'Permanência', *'Dropout'* foi mapeado para 'Evasão'. A Figura 1 mostra a distribuição após o remapeamento.

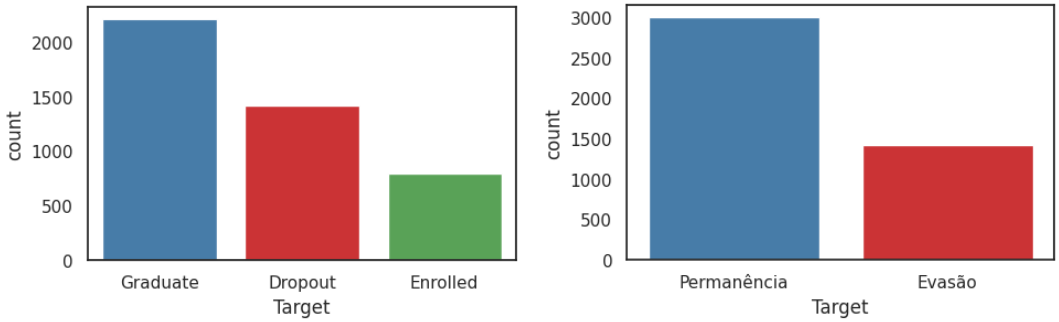


Figura 1 - Distribuição de frequência para as classes do atributo alvo, para as classes originais e para as novas classes reagrupadas.

P3. Quais os padrões e anomalias dos atributos individuais?

Para cada atributo do conjunto de dados, foram realizadas estatísticas descritivas em uma análise univariada. Para os atributos numéricos, calculou-se: frequência de cada classe, média, mediana, variância, desvio padrão, valor mínimo e máximo, intervalo de valores, número de amostras únicas e faltantes, Q1, Q3, Intervalo Interquartil (IQR), limite inferior e superior, obliquidade e curtose. Os gráficos de distribuição de frequência para todos os atributos estão no Anexo 2. A análise identificou que não existem valores ausentes ou potenciais erros nos dados.

Foram observadas instâncias discrepantes (*outliers*) em alguns atributos, como A*pplication order, Curricular units e Age at Enrollment*. No entanto, uma análise detalhada dos valores mínimos, máximos e limites revelou que grande parte dessas discrepâncias não correspondem a inconsistências ou erros. Para os dados relativos às unidades curriculares do primeiro e segundo semestre (*Curricular units*), encontrou-se pouca informação sobre a metodologia e organização destas métricas, dado o conhecimento limitado que se tem sobre a organização acadêmica da instituição para avaliar as disciplinas curriculares. Por exemplo, o número de disciplinas matriculadas por alunos no primeiro e segundo semestre tem valores máximos entre 23 e 26 unidades curriculares, sendo que a média é próxima de 6, como mostra a Figura 2. Uma possível hipótese para justificar os valores discrepantes é que o semestre se refere a um conjunto de disciplinas. Se o estudante continua matriculado nestas disciplinas repetidas vezes, são contabilizadas como unidades curriculares matriculadas no primeiro e segundo semestre.

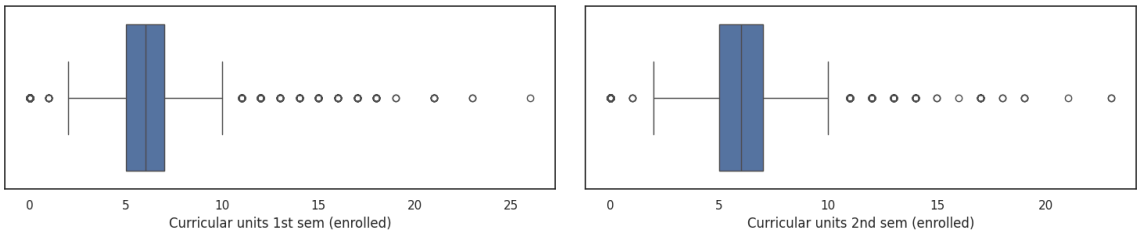


Figura 2 - *Box plot* de unidades curriculares matriculadas por alunos no primeiro e segundo semestre.

Dentre os atributos numéricos, as distribuições que parecem separar melhor os dados são relativas às notas e número de aprovações, tanto no primeiro quanto no segundo semestre, conforme as Figuras 3 e 4.

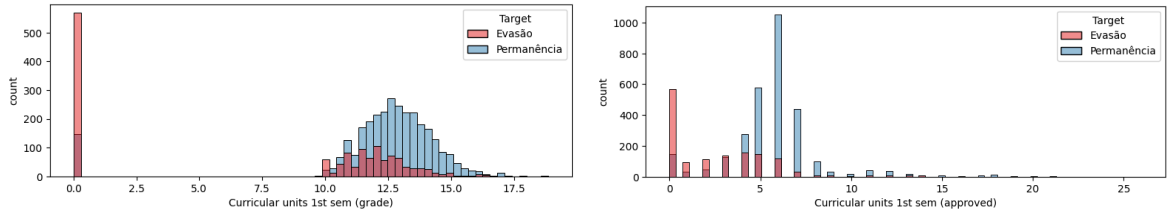


Figura 3 - Distribuição de frequência dos atributos relativos à nota e número de aprovações no 1o semestre.

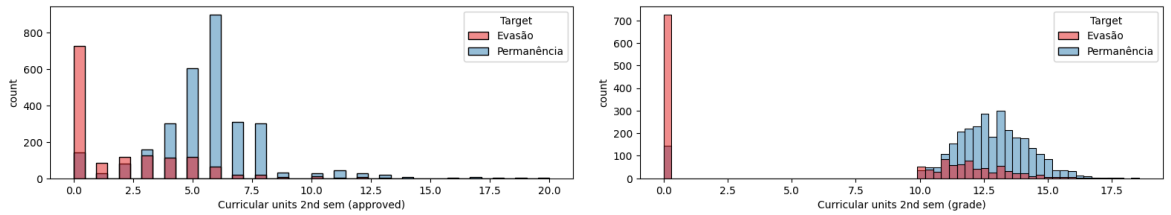


Figura 4 - Distribuição de frequência dos atributos relativos à nota e número de aprovações no 2o semestre.

Para os atributos categóricos, foram calculadas apenas duas estatísticas descritivas: a frequência de cada classe e a moda. Os gráficos de distribuição de frequência de todos os atributos estão disponíveis no Anexo 3. Observa-se que os atributos relacionados à qualificação e ocupação apresentam um grande desbalanceamento, com a maioria das classes possuindo menos de 50 instâncias. Além disso, destaca-se uma diferença significativa entre os gêneros binários, como mostra a Figura 5: as instâncias do gênero binário feminino demonstraram uma taxa de permanência significativamente maior, tanto em comparação com a evasão dentro do mesmo grupo, quanto em relação ao gênero binário masculino.

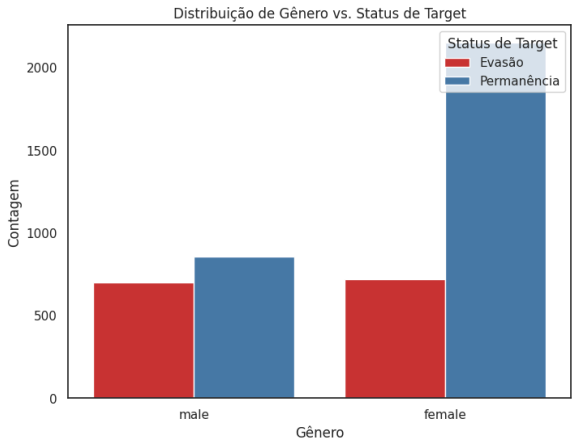


Figura 5 - Distribuição de frequência de evasão e permanência nos gêneros binários masculino e feminino.

Dos atributos categóricos, os que aparentam melhor separar os dados são os atributos binários relacionados ao pagamento em dia da mensalidade (*Tuition fees up to date*) e se o estudante é bolsista (*Scholarship holder*), conforme Figura 6. Estudantes com o pagamento atrasado tendem mais a evadir do que permanecer. Já os estudantes bolsistas tendem mais a permanecer do que evadir.

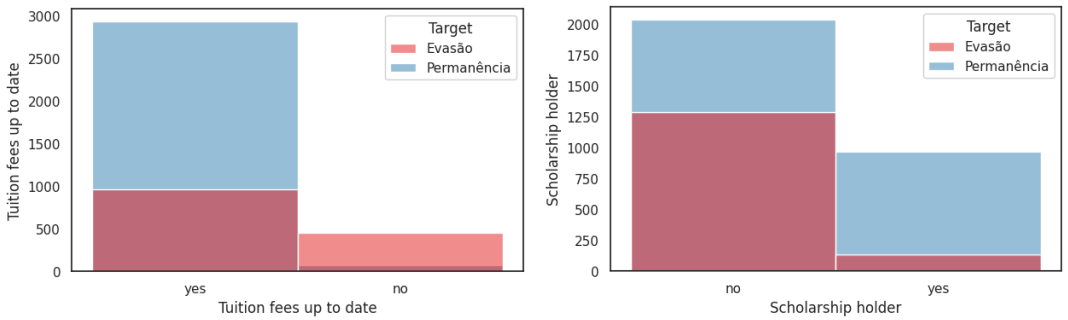


Figura 6 - Distribuição de frequência dos atributos relativos ao pagamento em dia e se o estudante é bolsista.

P4. Quais os padrões e anomalias entre os atributos?

Para comparar o conjunto de atributos, realizamos dois tipos de análises: bivariada, buscando relações entre pares de atributos; e multivariada, buscando relações entre subconjuntos de atributos. Na análise bivariada, buscamos investigar se existia separação das classes entre as notas do primeiro e do segundo semestre. Plotamos os dois atributos em um gráfico de dispersão, conforme Figura 7a, porém não foi visualizada uma separação significativa. Estudantes evadidos tendem a ter notas similares entre o primeiro e segundo semestre. Esta tendência se confirma no número de aprovações entre os dois primeiros semestres, apresentado na Figura 7b.

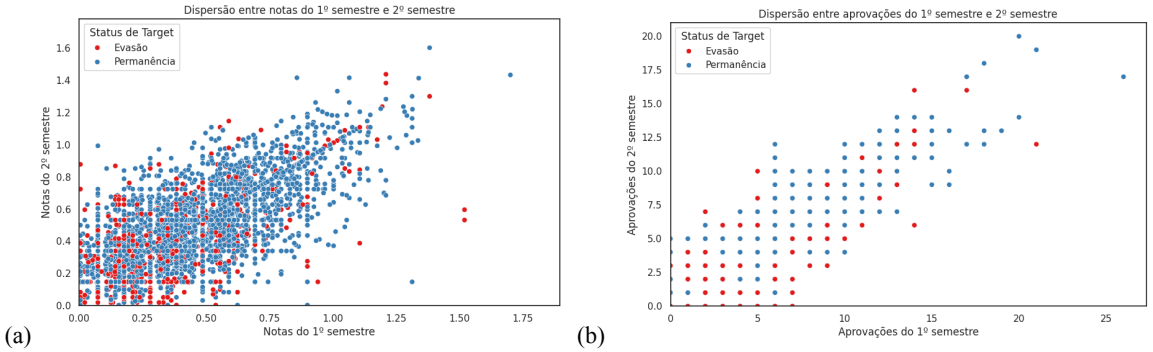


Figura 7 - (a) Gráfico de dispersão das notas dos alunos entre o 1o e 2o semestre. (b) Gráfico de dispersão do número de aprovações dos alunos entre o 1o e o 2o semestre.

Realizou-se uma análise de correlação de Pearson entre os atributos potencialmente preditivos e o atributo alvo, buscando avaliar quais fatores influenciam diretamente a desistência dos alunos. O vetor de correlação é apresentado na Figura 8.

Figura 8 - Correlação de Pearson entre cada atributo com o alvo (*Target*), com cor proporcional à correlação positiva (quente) ou negativa (frio) e intensidade da cor conforme a classificação convencional.

A classificação da correlação de Pearson pode variar entre diferentes autores e domínios. Uma possível abordagem convencional é mencionada por [3] e é apresentada na Tabela 1. Desta análise, 20 atributos apresentaram correlação desprezível, 9 correlação fraca e 5 correlação moderada com o atributo alvo, conforme classificado pela escala de cores da Figura 8.

Tabela 1 - Classificação usual de intervalos do coeficiente de correlação de Pearson [3].

Classificação	**Intervalo do coeficiente de correlação de Pearson (+/-)**
Muito forte	0.90 - 1.00
Forte	0.70 - 0.89
Moderada	0.40 - 0.69
Fraca	0.10 - 0.39
Desconsiderável	0.00 - 0.10

Em uma análise multivariada, calculamos a correlação cruzada entre todos os atributos do conjunto de dados e plotamos em uma matriz diagonal de correlação, conforme Figura 9.

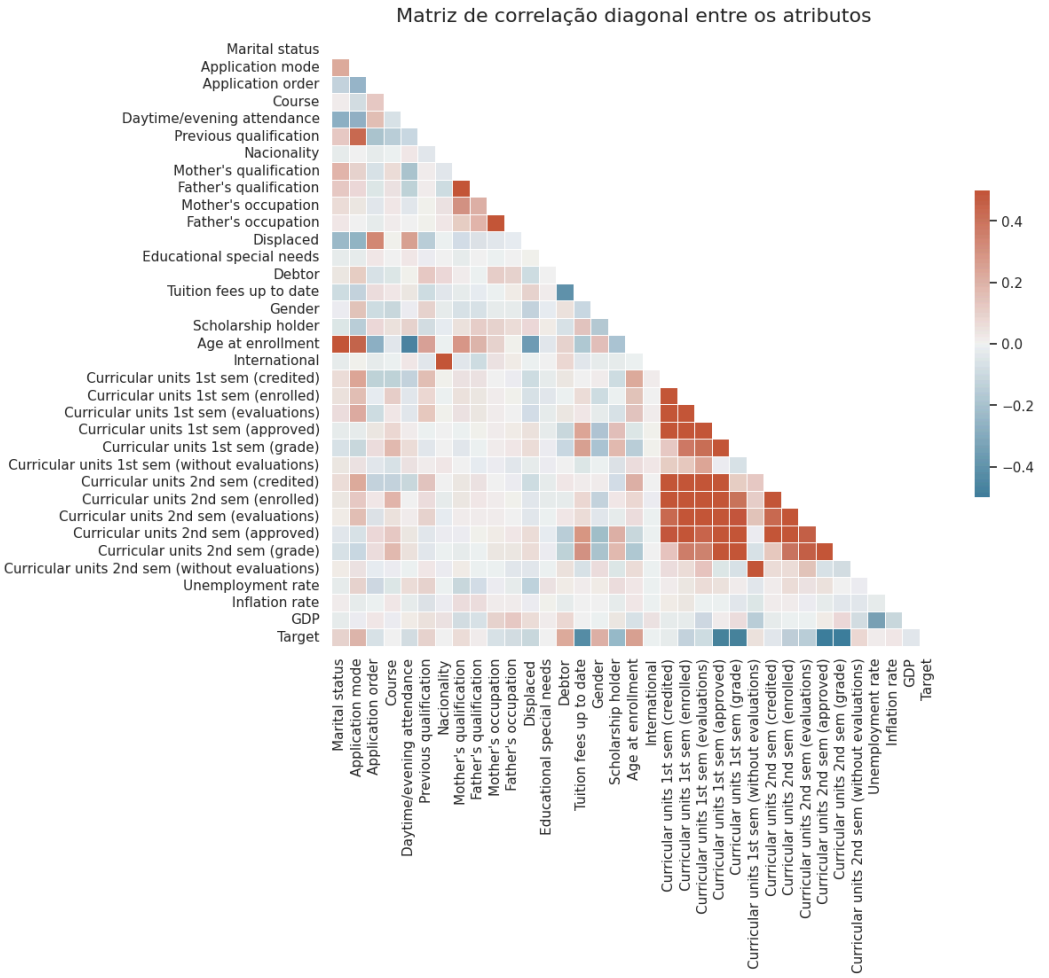


Figura 9 - Matriz diagonal da correlação de Pearson entre todos os pares de atributos, com cores divergentes e intensidade proporcionais à correlação positiva (quente) ou negativa (frio).

Como demonstra a Figura 9, existem muitos atributos com correlação significativa entre si, sendo potencialmente redundantes como atributos preditivos. Exemplos destes pares são:

- *Age at enrollment / Marital status*
- *Age at enrollment / Application mode*
- *Mother's occupation / Father's occupation*
- *Mother's qualification / Father's qualification*
- *International / Nacionality*
- *Curricular units 1st sem / Curricular units 2nd sem*

Portanto, removemos um atributo destes pares identificados com correlação forte ou moderada entre estes. Os atributos removidos foram: '*Age at enrollment*', '*Father's occupation*', '*Father's qualification*', "*Nacionality*", '*Curricular units 1st sem (grade)*', '*Curricular units 1st sem (without evaluations)*', assim como todas relacionadas a '*enrolled*', '*evaluations*', '*credited*' e '*approved*' de ambos os semestres. Manteve-se '*Curricular units 2nd sem (grade)*', '*Curricular units 2nd sem (without evaluations)*' para representar os atributos relativos a dados curriculares. As Figuras 10 e 11 indicam o efeito da remoção destes atributos na análise de correlação uni e multivariada.

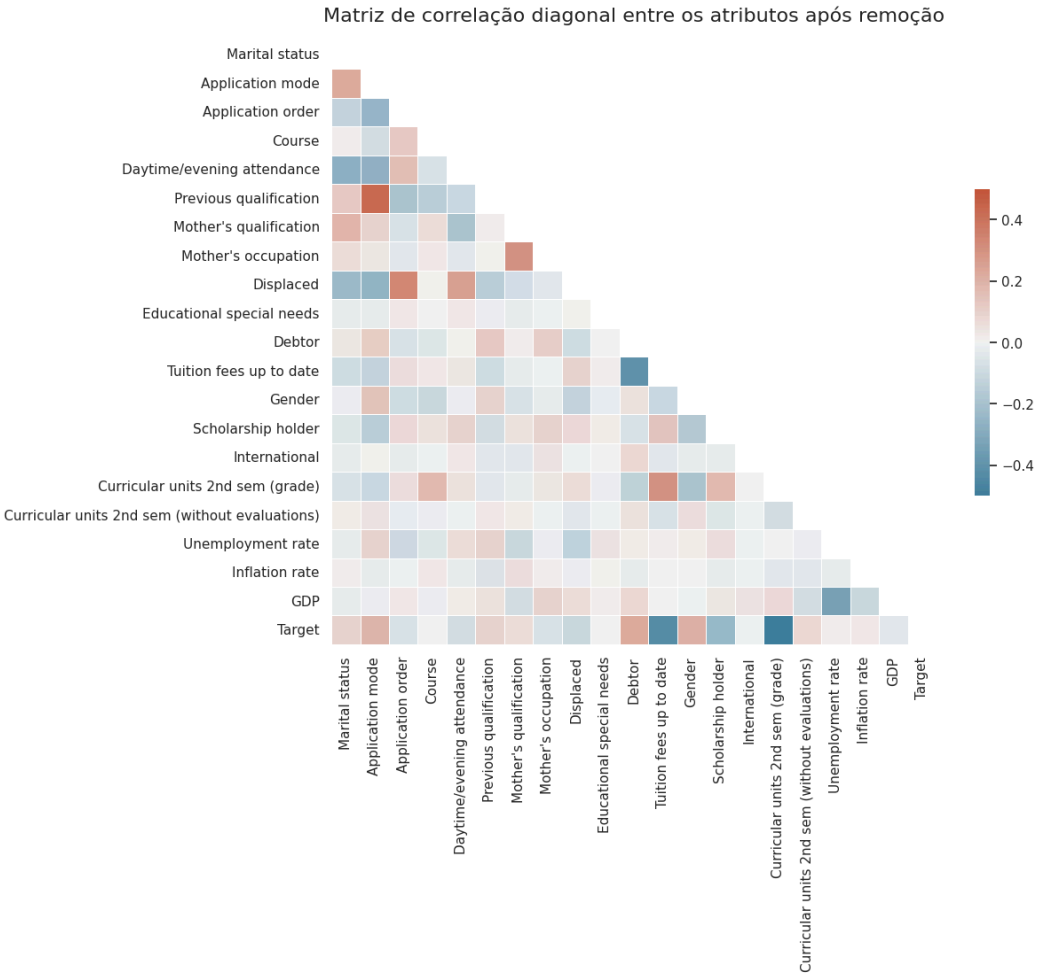


Figura 10 - Matriz diagonal da correlação de Pearson entre todos os pares de atributos, após remoção de atributos com correlação forte ou moderada entre si. Apresenta cores divergentes e intensidade proporcionais à correlação positiva (quente) ou negativa (frio).

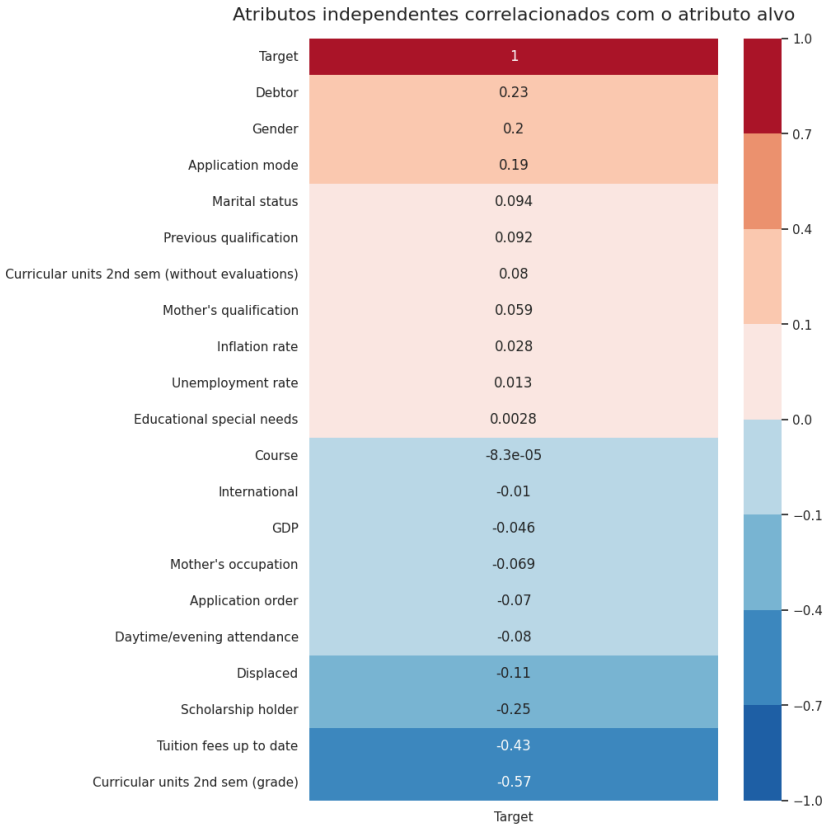


Figura 11 - Correlação de Pearson entre cada atributo com o alvo (*Target*), após remoção de atributos com correlação forte ou moderada entre si com cor proporcional à correlação positiva (quente) ou negativa (frio) e intensidade da cor conforme a classificação convencional.

2.2 Pré-processamento dos Dados

Durante a análise exploratória, conhecemos os dados e identificamos padrões, anomalias e potenciais atributos preditivos. Inicialmente, observamos que todos os 34 atributos potencialmente treináveis estavam codificados como valores numéricos (*float64* ou *int64*), embora alguns fossem categóricos. Isso dificultava a interpretação e a análise entre atributos categóricos e numéricos. Para solucionar, reconstruímos os atributos categóricos em um novo *data frame* com base no descritor de dados do *dataset* [1]. Após a interpretação na análise exploratória, utilizamos o conjunto de dados já codificado e realizamos a codificação do atributo *Target* para um atributo numérico binário, com objetivo de utilizá-lo nos modelos preditivos.

A realização de etapas de pré-processamento a partir de todo o conjunto de dados pode ser uma fonte de "vazamento" de dados de treinamento para os dados de teste ou validação, pois neste caso o pré-processamento utiliza informações que representam todo o conjunto de dados para transformá-los. Ou seja, os dados particionados posteriormente para teste ou validação já terão sofrido influência prévia dos dados particionados para treino, podendo causar sobreajuste e afetar a metodologia de avaliação. Queremos supor que os conjuntos de validação e de teste são dados nunca antes vistos antes pelo modelo ajustado aos dados de treinamento, ou seja, independentes.

Para evitar que este tipo de "vazamento" aconteça, utilizamos a classe *Pipeline* da biblioteca *Scikit-learn*. Uma *pipeline* de modelagem se trata da sequência de métodos de transformação de

dados, conjunto de hiperparâmetros e algoritmo preditivo [4]. A classe provê uma sequência de métodos transformadores de dados, que permitem aplicar uma sequência de transformações exclusivamente sobre os dados de treinamento, isolando-as dos dados de teste e evitando o "vazamento" de dados. Também oferece um método preditor, para aplicar a *pipeline* após seu treinamento aos dados de teste.

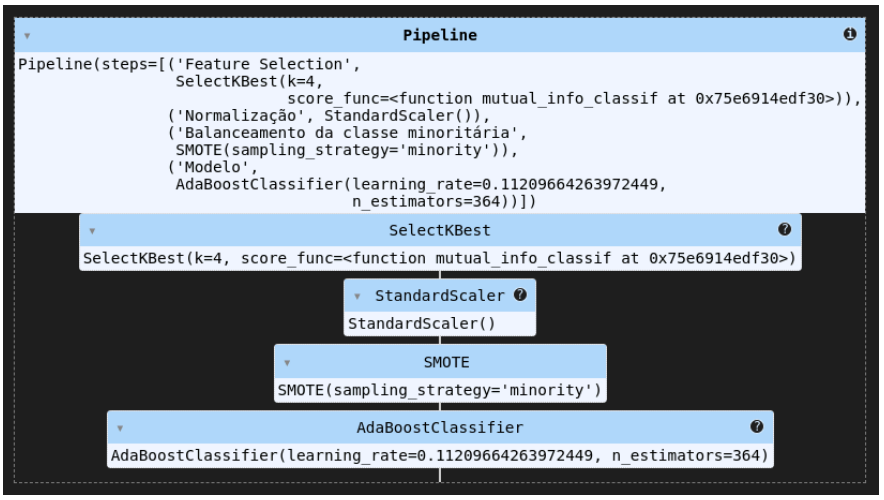


Figura 12 - *Pipeline* de modelagem adotada, onde o *AdaBoostClassifier* representa um exemplo de um dos possíveis algoritmos e conjunto de hiperparâmetros a serem utilizados.

A seleção de atributos é realizada através da classe *SelectKBest* do *scikit-learn*, que seleciona os *k* atributos com maior pontuação, resultado de uma função de pontuação [5]. A função de pontuação utilizada foi *mutual_info_classif*, a qual estima a informação mútua para uma variável alvo discreta, utilizando métodos não-paramétricos baseados na estimação de entropia entre distâncias de *k-nearest neighbors* [6]. A informação mútua mede, portanto, a dependência entre duas variáveis, sendo indicada quando o dataset é composto por variáveis tanto numéricas quanto categóricas [7]. Selecionamos empiricamente um valor de *k=4* para o *SelectKBest,* pois verificamos com os resultados do *Spot-checking* que era um número suficiente dado a importância dos atributos para a predição, o que será aprofundado na seção 7.

Dentre os atributos selecionados, nenhum se trata de um atributo categórico nominal. Por isso, não foi necessária a realização de transformação via *one-hot encoding*. Experimentou-se a transformação de atributos categóricos nominais codificados para novos atributos com *one-hot encoding*, para atributos não selecionados na Análise Exploratória. O objetivo foi explorar se os novos atributos derivados poderiam modificar as relações de correlação. Todavia, não houve impacto significativo para a mudança dos atributos selecionados.

Para a normalização dos dados adicionamos o *StandardScaler* do *scikit-learn* como etapa da *pipeline*, pois cálculo de *z-score* depende da média e do desvio padrão de todo o conjunto de dados [8]. Assim, garante-se que os dados de teste não sejam contaminados por informações dos dados de treinamento ao serem normalizados.

O atributo alvo apresentava um alto desbalanceamento entre as classes, detectado na Análise Exploratória e representado pela Figura 1. Para mitigar este problema, foi adicionada na *pipeline* uma etapa de balanceamento de classes através do método *SMOTE* com estratégia minoritária, implementado na biblioteca *imbalanced-learn* [9].

3. Definição da Abordagem, Algoritmos e Estratégias de Avaliação

3.1 Seleção dos Algoritmos

Selecionamos nove algoritmos de aprendizado supervisionado para classificação:

Tabela 3 - Algoritmos e hiperparâmetros selecionados para o processo de *spot-checking*.

Algoritmo de aprendizado	**Hiperparâmetros**	**Justificativa**
Árvore de Decisão	● Critério: gini ● Sem poda	Simples interpretação e visualização das decisões.
Árvore de Decisão	● Critério: gini ● Altura máxima de 5	Variação para entender o impacto do uso de poda.
Floresta Aleatória	● Critério: gini	Combinação de múltiplas árvores para melhorar a precisão e reduzir o *overfitting*.
Floresta Aleatória	● Critério: entropia	Variação para entender o impacto do critério informativo.
Regressão Logística Multinomial	● Padrão	Modelo linear adequado para classificação multiclasse.
K-Nearest Neighbors	● K = 3	Classificação baseada na proximidade das instâncias nos dados.
K-Nearest Neighbors	● K = 5	Variação para entender o impacto de adicionar mais vizinhos.
Máquina de Vetores de Suporte	● Kernel linear	Busca maximizar a margem entre as classes.
AdaBoost	● No. de estimadores = 50 ● Taxa de aprendizado = 1 ● Algoritmo: SAMME	Ajusta classificações anteriores em favor das instâncias classificadas negativamente.

Os algoritmos selecionados abrangem diferentes vieses indutivos, permitindo uma comparação abrangente entre algoritmos de classificação de forma rápida, por meio do processo de *spot-checking*. Além disso, introduzimos variações em hiperparâmetros de alguns algoritmos vistos em aula, para validar seu desempenho no conjunto de dados utilizado.

3.2 Métricas de Desempenho

Selecionamos as seguintes métricas:

- **F1-score**: média harmônica entre precisão e revocação.
- **Precisão:** fração das predições positivas que estão corretas.
- **Revocação:** avalia a taxa de acerto na classe positiva.

- **ROC AUC**: avalia a performance de um classificador entre diferentes valores de *threshold*, através do *tradeoff* entre sensibilidade e especificidade.

Métrica Principal: **F1-Score** foi escolhida como critério principal de seleção dos modelos, pois como o balanceamento acontece no *pipeline*, os dados de teste se encontram desbalanceados. Como esta métrica considera juntos a precisão e a revocação, o *F1-score* é indicado como métrica de avaliação de modelos para dados desbalanceados ao invés de acurácia, que pode suprimir erros para classes minoritárias.

3.3 Estratégia de Avaliação

Utilizou-se a estratégia de avaliação cruzada *K-Fold* repetida, variando entre cinco sementes aleatórias diferentes, para levar em conta diferentes variações de desempenho devido ao particionamento mas garantindo reprodutibilidade. A estratégia de validação cruzada é importante para que todos os dados possam ser utilizados na avaliação, garantindo sua robustez. Utilizou-se a implementação *cross-validate* do *scikit-learn*, pois possibilita o cálculo de várias métricas em uma mesma chamada do método.

Utilizou-se a *pipeline* de modelagem citada na subseção 2.2 para cada novo *fold* da validação cruzada, evitando assim a contaminação de dados.

4. Spot-Checking de Algoritmos

4.1 Estratégia de Spot-Checking

Realizou-se uma avaliação cruzada 10-*fold* para cada algoritmo, passando pela *pipeline* pré-definida. Ao retornar as métricas, foi realizado seu pós-processamento em um *data frame* que as relaciona com o número da iteração de *fold* e o algoritmo que foi avaliado. Assim, obteve-se o resultado completo da avaliação no *metrics_df*, o que permitiu sua sumarização com estatísticas descritivas como média e desvio padrão, além da construção de gráficos de *box plot.*

Divisão dos Dados:

- **Características (X)**: todos os atributos preditores.
    - **Antes do pré-processamento:** *Marital status, Application mode, Application order, Course, Daytime/evening attendance, Previous qualification, Mother's qualification, Mother's occupation, Displaced, Educational special needs, Debtor, Tuition fees up to date, Gender, Scholarship holder, International, Curricular units 2nd sem (grade), Curricular units 2nd sem (without evaluations), Unemployment rate, Inflation rate, GDP*.
    - **Após seleção de atributos *Select4Best*:** *Application mode, Tuition fees up to date, Scholarship holder, Curricular units 2nd sem (grade).*
- **Alvo (y)**: atributo "*Target*" binário, onde 0 significa 'Permanência' e 1 significa 'Evasão'.

4.2 Resultados

A Tabela 4 apresenta a sumarização dos resultados médios para cada métrica e algoritmo, acompanhados de seu desvio padrão. Em negrito, destaca-se os três algoritmos com a melhor métrica.

Tabela 4 - Resultado da validação cruzada 10-*fold*. Cada métrica é representada por sua média e desvio padrão.

Algoritmo	**F1-Score**	**Precisão**	**Revocação**	**ROC AUC**
Árvore de Decisão	0.683 (0.029)	0.658 (0.041)	0.713 (0.044)	0.783 (0.024)
Árvore de Decisão (Max Depth 5)	0.703 (0.033)	0.703 (0.04)	0.705 (0.05)	0.825 (0.029)
Floresta Aleatória (Gini)	0.687 (0.032)	0.654 (0.042)	**0.727 (0.044)**	0.843 (0.021)
Floresta Aleatória (Entropia)	0.689 (0.035)	0.656 (0.044)	**0.728 (0.047)**	0.846 (0.022)
Regressão Logística	**0.734 (0.03)**	**0.784 (0.04)**	0.692 (0.043)	**0.862 (0.022)**
3-*Nearest Neighbors*	0.679 (0.04)	0.678 (0.044)	0.682 (0.054)	0.808 (0.026)
5-*Nearest Neighbors*	0.692 (0.033)	0.692 (0.045)	0.697 (0.053)	0.832 (0.021)
AdaBoost	**0.733 (0.029)**	**0.726 (0.037)**	**0.743 (0.046)**	**0.871 (0.019)**
Máquina de Vetores de Suporte	**0.726 (0.031)**	**0.818 (0.033)**	0.653 (0.041)	**0.863 (0.021)**

A Figura 13 sumariza as métricas obtidas na avaliação do *spot-checking,* divididas entre F1 *score*, precisão, revocação e área sob a curva ROC.

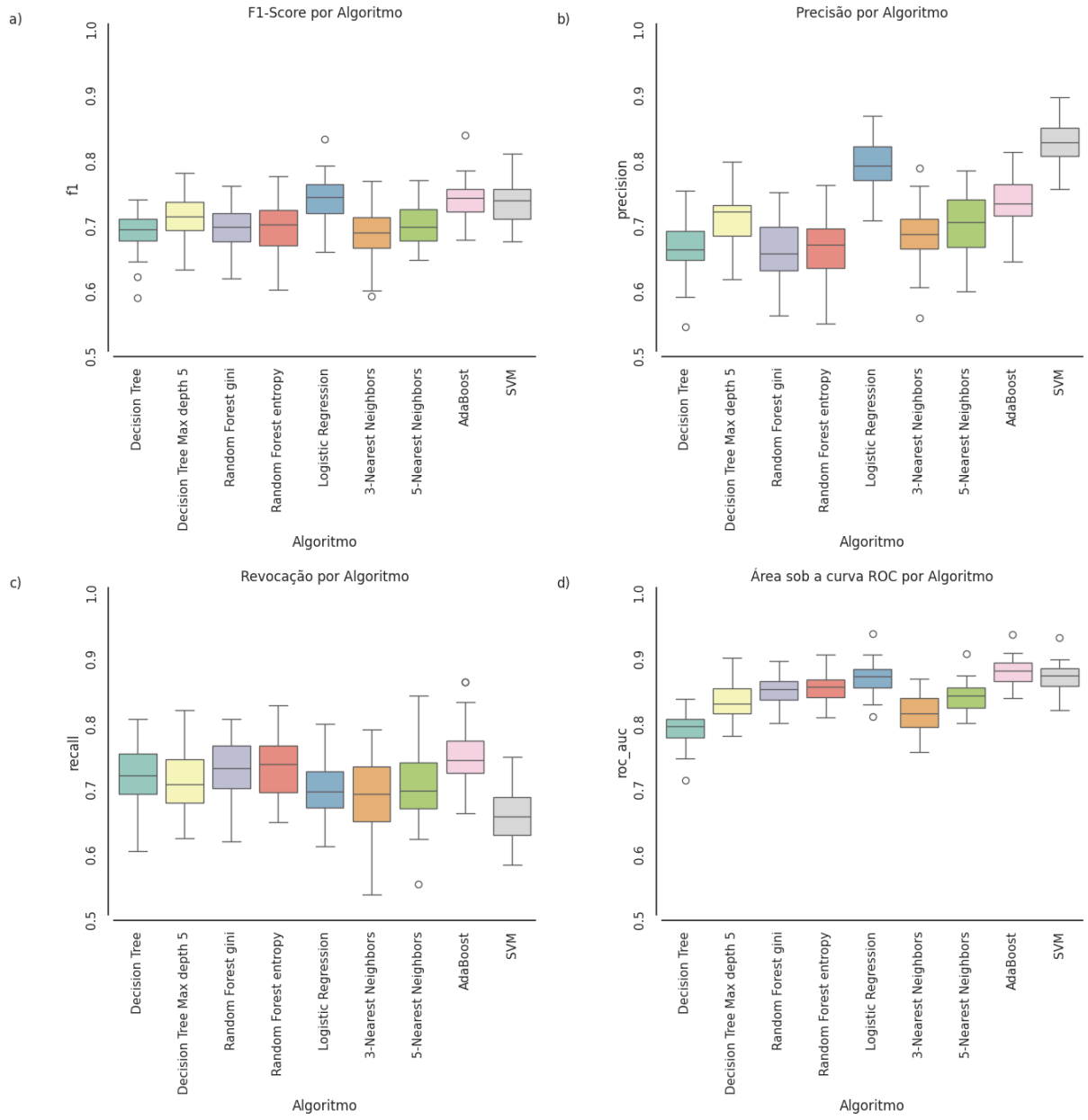


Figura 13 - Resultados do *Spot-checking* com validação cruzada 10-*fold*. (a) *Box plot* do F1-score por algoritmo. (b) *Box plot* da precisão por algoritmo. (c) *Box plot* da revocação por algoritmo. (d) *Box plot* do ROC AUC por algoritmo.

A partir dos resultados obtidos na avaliação do *Spot-checking*, selecionamos três algoritmos com o melhor *F1-Score*, conforme indicado na Tabela 4 e Figura 13a para a etapa de otimização de hiperparâmetros: Regressão Logística, Máquina de Vetores de Suporte e AdaBoost.

5. Otimização de hiperparâmetros

Visando obter o algoritmo com o melhor desempenho possível, realizou-se um processo de otimização de hiperparâmetros para os algoritmos selecionados no *spot-checking*. Para isto, foi utilizada a biblioteca Optuna, um *framework* de otimização automática de hiperparâmetros para aprendizado de máquina. A biblioteca oferece algoritmos do estado da arte para a amostragem de hiperparâmetros, espaços de busca com sintaxe *Pythonica* e oferece visualizações rápidas.

Como estratégia de divisão de dados, realizou-se uma partição de dados 70-30, onde 70% dos dados foram utilizados para a otimização de hiperparâmetros, e 30% dos dados foram reservados para teste, a serem avaliados na seção 6. Para esta primeira partição, aplicou-se um processo de validação cruzada aninhada, pois garante uma maior variabilidade nos dados utilizados para otimização de hiperparâmetros e sua avaliação. A partir da estimativa de desempenho de cada conjunto de hiperparâmetros no conjunto de teste, foram selecionados os melhores hiperparâmetros para cada algoritmo através da métrica principal *F1-score.*

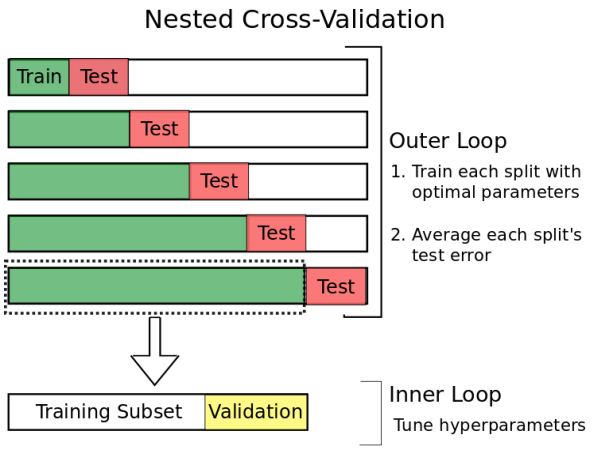


Figura 14 - Estratégia de divisão de dados Validação Cruzada Aninhada. Fonte: https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9

No Optuna, um estudo se refere a uma otimização completa baseada em uma função objetivo. Já uma tentativa se trata de uma execução única de uma função objetivo. Neste trabalho, a função objetivo implementada é composta pela sequência entre definição do espaço de otimização, criação da pipeline com parâmetros escolhidos por uma tentativa do Optuna, validação cruzada interna com $k=5$ *folds* e métrica de avaliação média da validação. Portanto, cada *fold* da validação cruzada externa executa um estudo, que é composto por um determinado número de tentativas, e retorna os hiperparâmetros avaliados no conjunto de validação interno através da métrica principal *F1-score*, onde os melhores hiperparâmetros são avaliados no conjunto de teste externo. As Figuras 14 e 15 ilustram os processos de validação cruzada aninhada e otimização de hiperparâmetros, aplicada na validação cruzada interna.

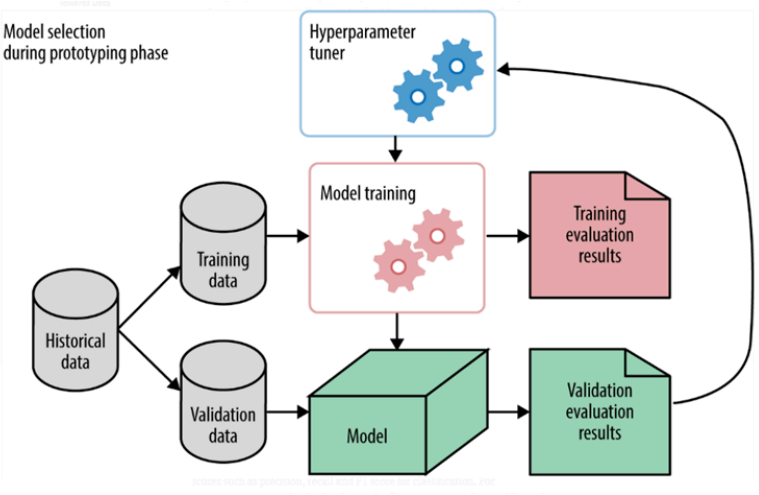


Figura 15 - Representação do uso do Optuna para otimização de hiperparâmetros e seleção de modelos, aplicado na validação cruzada interna. Fonte: https://medium.com/@kalyaniavhale7/understanding-of-optuna-a-machine-learning-hyperparameter-optimization-framework-ed31ebb335b9

A definição do espaço de otimização foi realizada empiricamente para cada algoritmo selecionado. A Tabela 5 apresenta, para cada algoritmo, os respectivos hiperparâmetros a serem otimizados, assim como seu tipo e seu espaço de busca.

Tabela 5 - Espaço de busca de hiperparâmetros para cada algoritmo.

Algoritmo	**Hiperparâmetro**	**Tipo**	**Espaço de busca**
Regressão Logística	C	float (log)	$[10^{-3}, 10^{3}]$
	solver_penalty	categorical	[liblinear_l1, liblinear_l2, lbfgs_l2]
AdaBoost	n_estimators	int	[50, 500]
	learning_rate	float (log)	[0.01, 2.00]
Máquina de Vetores de Suporte	C	float (log)	$[10^{-4}, 10^{2}]$
	kernel	categorical	[linear, rbf, sigmoid]

5.1 Regressão Logística

Para o algoritmo de Regressão Logística, os hiperparâmetros a serem otimizados foram os hiperparâmetros C e penalidade do *solver*.

O hiperparâmetro C se trata do inverso do fator de regularização, onde valores pequenos especificam uma regularização maior [10].

A penalidade se trata do método de regularização utilizado para prevenir sobreajuste ao adicionar um termo de penalidade na função de custo. A regularização L1 adiciona o valor absoluto de magnitude do coeficiente como termo de penalidade para a função de custo. Se o atributo não está contribuindo de forma significativa para a Regressão Logística, o L1 penaliza os pesos para este atributo para um valor próximo a zero. A regularização L2 adiciona um termo que é proporcional à soma dos quadrados dos coeficientes do modelo, desencorajando-o a atribuir muita importância a um único atributo [11]. Os termos 'liblinear' e 'lbfgs' se referem às bibliotecas que implementam os *solvers* das regularizações utilizando métodos distintos.

A Figura 16 apresenta o histórico do melhor estudo na otimização destes hiperparâmetros para a Regressão Logística, onde cada valor objetivo é o resultado de uma função objetivo. A Figura 17 apresenta os valores de hiperparâmetro para cada valor objetivo.

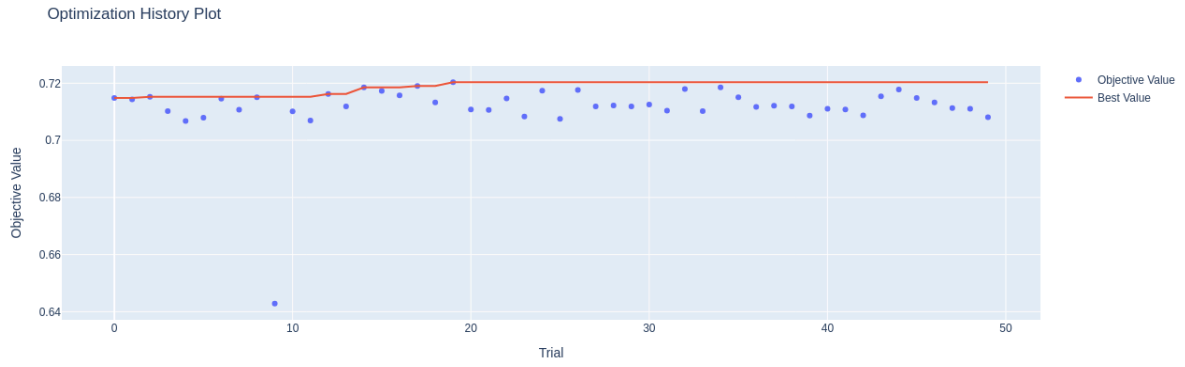


Figura 16 - Histórico do melhor estudo para a otimização de hiperparâmetros da Regressão Logística.

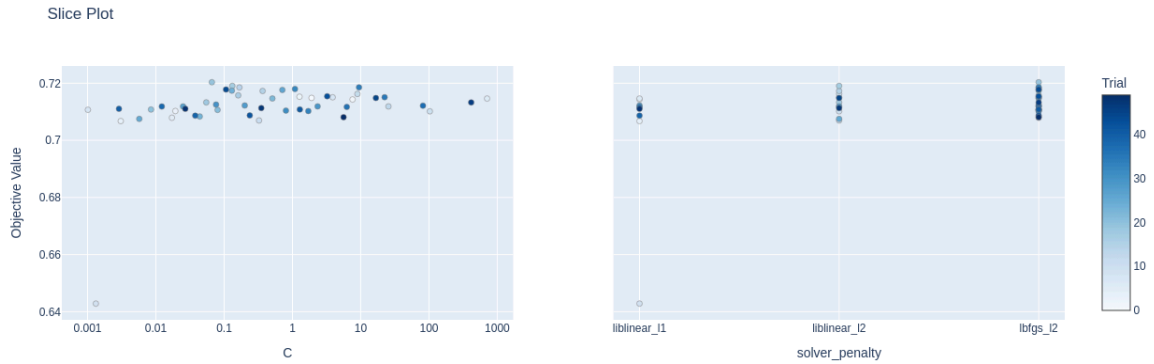


Figura 17 - Valores-objetivo para cada hiperparâmetro na otimização da Regressão Logística.

5.2 AdaBoost

Para o algoritmo classificador AdaBoost, os hiperparâmetros a serem otimizados foram o número de estimadores e a taxa de aprendizado.

O número de estimadores é utilizado como uma condição de parada do processo de *boosting*, determinando um número máximo de estimadores construídos pelo algoritmo. Já a taxa de aprendizado é o peso aplicado a cada classificador em cada iteração de *boosting*, responsável por regular a contribuição de cada um. Existe uma relação de compromisso entre o número de estimadores e a taxa de aprendizado, pois se a taxa de aprendizado é muito baixa, os estimadores existentes estão contribuindo pouco para o *boosting*, possibilitando a criação de mais estimadores [12].

A Figura 18 apresenta o histórico do melhor estudo na otimização destes hiperparâmetros para o AdaBoosting, onde cada valor objetivo é o resultado de uma função objetivo. A Figura 19 apresenta os valores de hiperparâmetro para cada valor objetivo.

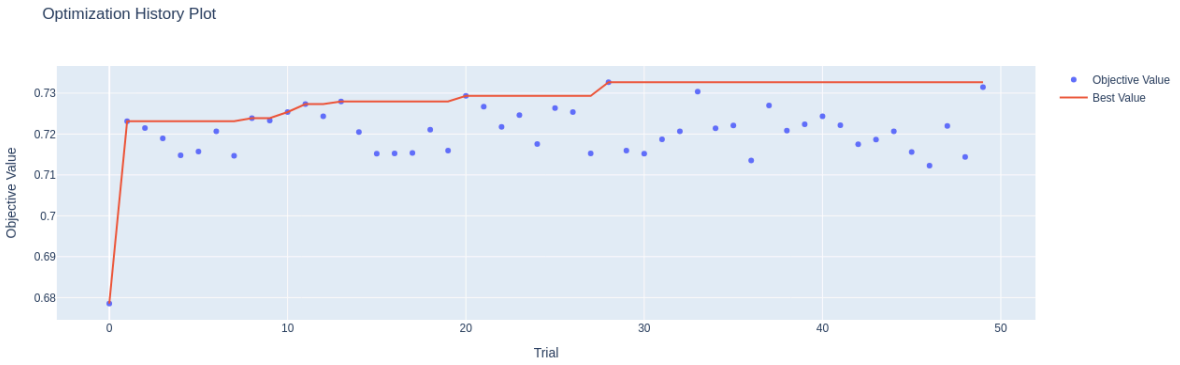


Figura 18 - Histórico do melhor estudo para a otimização de hiperparâmetros do AdaBoost.

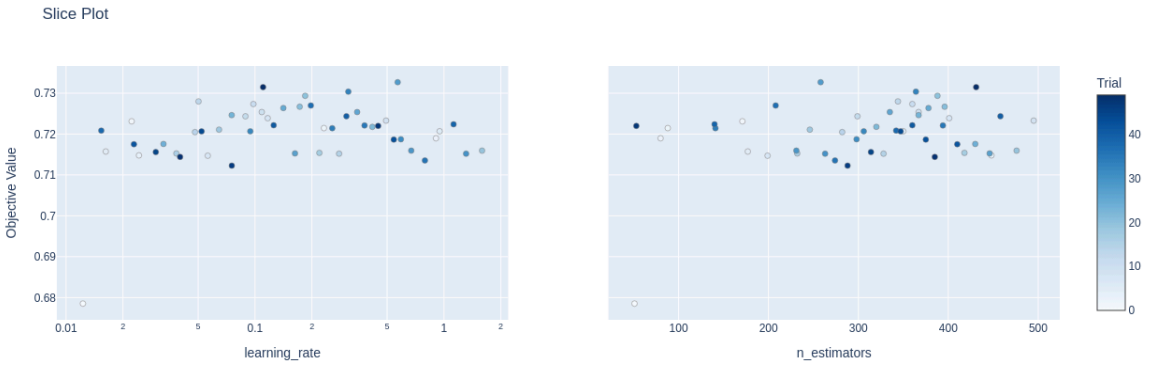


Figura 19 - Valores-objetivo para cada hiperparâmetro na otimização do AdaBoost.

5.3 SVM

Para o algoritmo Máquina de Vetores de Suporte (SVM), os hiperparâmetros a serem otimizados foram a função de *kernel* e o hiperparâmetro C

A função de *kernel* em SVMs é utilizada para aumentar a dimensionalidade do espaço de entrada. Um classificador não-linear, quando transformado para um espaço de alta dimensionalidade, pode ser visto como um classificador linear [13]. O *kernel* linear é representado por $< x, x' >$ e é utilizado quando os dados são linearmente separáveis [14]. O *kernel rbf* (*Radial Basis Function* ou *kernel* Gaussiano) é expresso por $exp(-\gamma ||x - x'||^2)$, onde γ define a influência de um único exemplo de treinamento, e mede a similaridade entre dados com base na sua distância euclidiana no espaço de entrada [13]. O *kernel* sigmoidal é expresso por $tanh(\gamma < x, x' > + r)$, sendo equivalente a um perceptron de duas camadas em uma rede neural, pois também utiliza uma função sigmoide como função de ativação [15].

O hiperparâmetro C se trata do inverso do fator de regularização, onde valores pequenos especificam uma regularização maior [16].

A Figura 20 apresenta o histórico do melhor estudo na otimização destes hiperparâmetros para o SVM, onde cada valor objetivo é o resultado de uma função objetivo. A Figura 21 apresenta os valores de hiperparâmetro para cada valor objetivo.

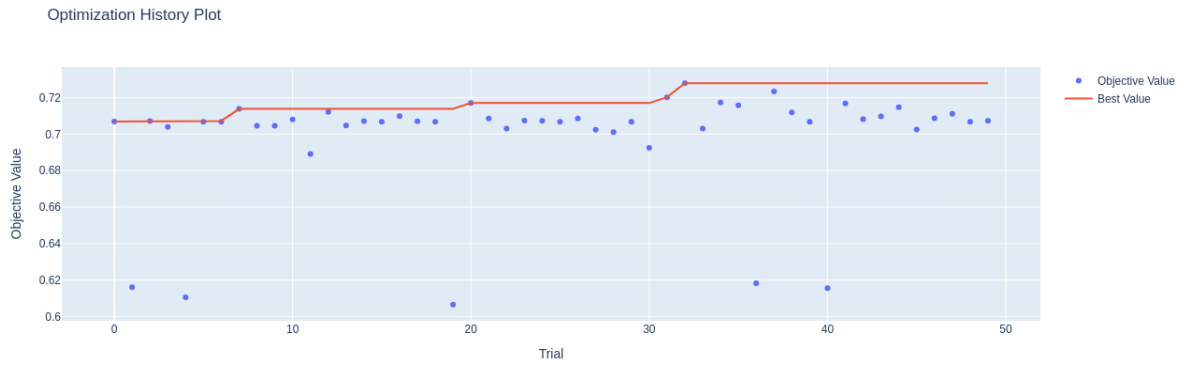


Figura 20 - Histórico do melhor estudo para a otimização de hiperparâmetros do SVM.

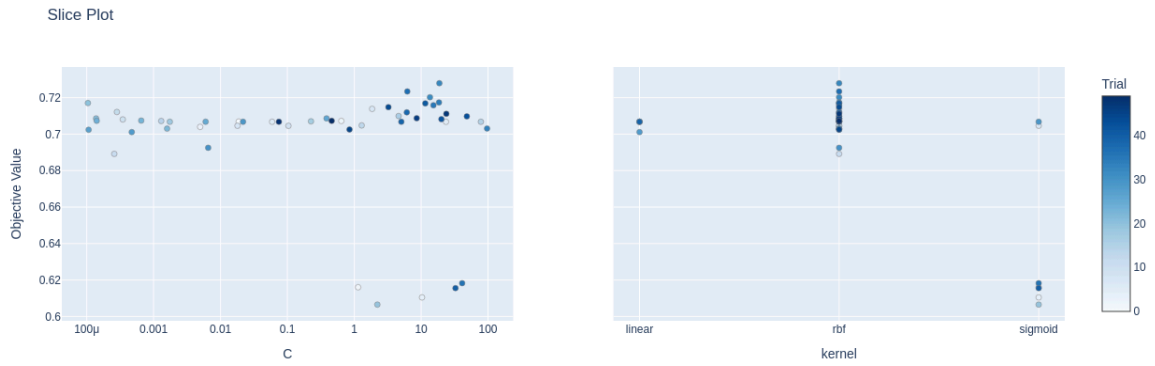


Figura 21 - Valores-objetivo para cada hiperparâmetro na otimização do SVM.

5.4 Avaliação de desempenho

Para avaliar os melhores parâmetros obtidos no processo de otimização, a Figura 22 sumariza a estimativa de desempenho nos 10 *folds* externos para cada algoritmo, entre *F1-Score* (Figura 22a), precisão (Figura 22b), revocação (Figura 22c) e área sob a curva ROC (Figura 22d).

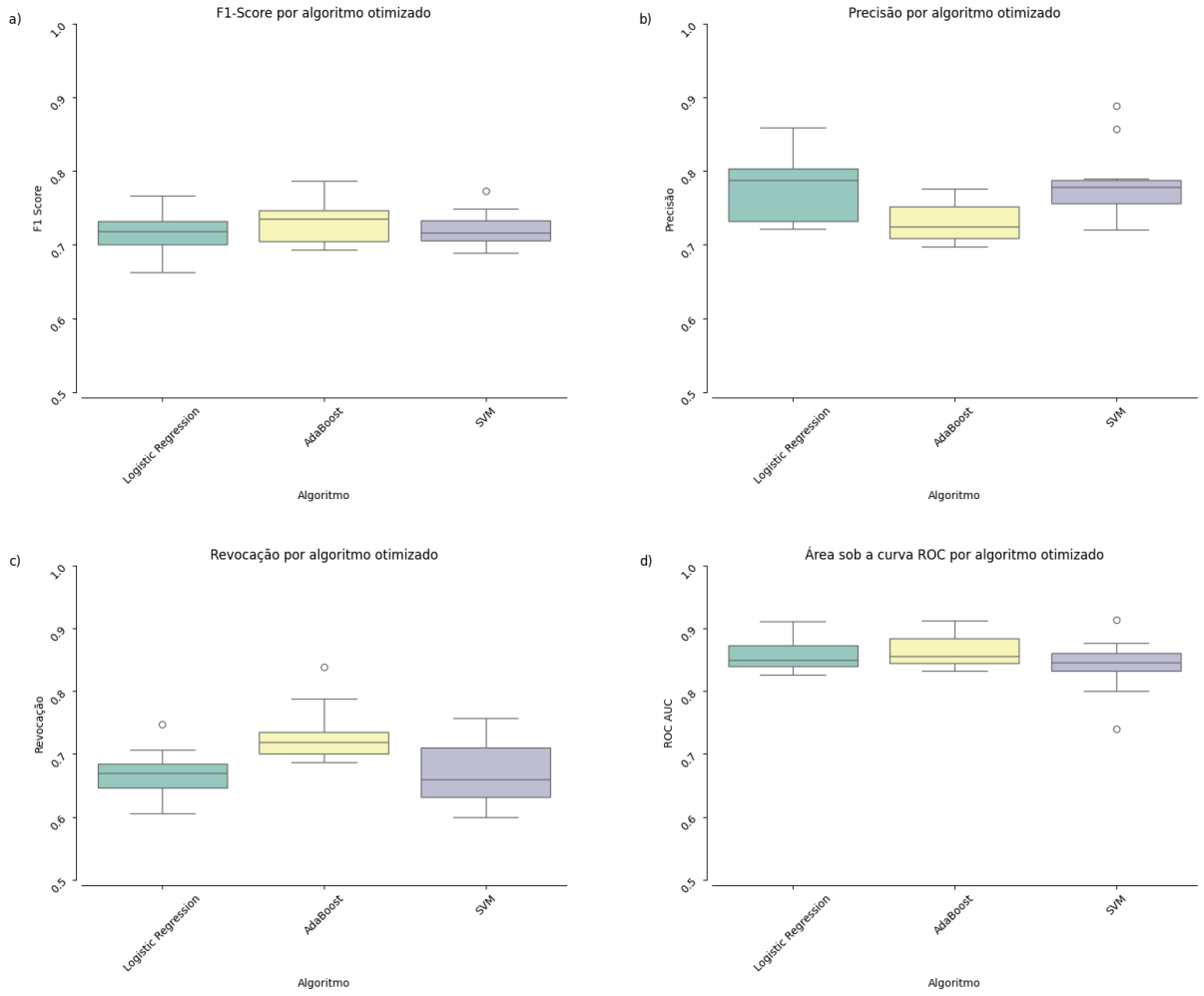


Figura 22 - Avaliação de desempenho dos algoritmos no processo de otimização de hiperparâmetros, para a validação cruzada externa de 10 *folds*. a) *Box-plot* do *F1-Score* por algoritmo otimizado. b) *Box-plot* da precisão por algoritmo otimizado. c) *Box-plot* da revocação por algoritmo otimizado. d) *Box-plot* da área sob a curva ROC por algoritmo otimizado.

Observa-se na Figura 22a que a distribuição de *F1-Score* para os três algoritmos é bastante similar, variando aproximadamente acima de 0,65 e abaixo de 0,80, com média entre 0,70 e 0,75. Isto expressa que a relação de compromisso entre precisão e revocação dos três algoritmos, com otimização de seus hiperparâmetros, é bastante similar. O algoritmo SVM apresenta o menor intervalo interquartil e amplitude da distribuição de *F1-Score*, mas por uma margem suficientemente pequena.

Observando a distribuição de precisão, apresentada na Figura 22b, observa-se que a Regressão Logística apresenta o maior valor médio de precisão dentre os três algoritmos, próximo de 80%, assim como a maior variância observada pela amplitude e intervalo interquartil. O SVM apresenta uma média próxima da Regressão Logística, enquanto AdaBoost apresenta uma média mais

próxima de 70%. Ambos têm uma amplitude menor, porém o SVM apresenta dois valores discrepantes, com performance acima de 80%.

Em termos de revocação, como apresentado na Figura 22c, a maior média é apresentada pelo AdaBoost, pouco acima de 70%, mas a distribuição de maior amplitude é do SVM, este que também tem a menor média. Portanto, ao comparar a revocação com a precisão, um menor desempenho em revocação é compensado em uma relação de compromisso por um melhor desempenho em precisão, e vice-versa, o que é sumarizado pelo *F1-Score*.

Em termos da área sob a curva ROC na figura 22d, os três algoritmos têm desempenho comparável, tanto em termos de média quanto em termo da distribuição de seus quartis. O algoritmo SVM apresenta ligeiramente a menor média, mas também apresenta dois valores discrepantes, um aproximadamente alinhado com o limite superior da distribuição dos dois outros algoritmos, e um abaixo dos limites inferiores da distribuição dos outros algoritmos.

O processo aplicado de otimização de hiperparâmetros mostrou que, para este conjunto de dados e seus atributos, assim como os vieses indutivos dos algoritmos selecionados, poucas tentativas foram necessárias para se encontrar valores ótimos de hiperparâmetros, os quais foram empregados para maximizar seu desempenho de acordo com as métricas selecionadas. A seção 7 sobre interpretação do melhor modelo terá o papel de investigar possíveis causas deste possível limite de otimização encontrado experimentalmente.

Por fim, selecionou-se o conjunto de hiperparâmetros que apresentou o melhor *F1-Score* no processo de otimização, sumarizados pela Tabela 6.

Tabela 6 - Hiperparâmetros selecionados pelo melhor *F1-Score* por algoritmo.

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 de desempenho em dados de teste

Anterior ao processo de otimização de hiperparâmetros, foi realizado um particionamento dos dados, onde uma partição de treino de 70% foi utilizada para otimização de hiperparâmetros. A partição reservada de treino, que representa 30% do conjunto de dados, foi utilizada para realizar uma avaliação final, voltada para a seleção do melhor modelo. Portanto, cada algoritmo foi treinado na partição de treino, utilizando os melhores hiperparâmetros selecionados na etapa anterior, e validado na partição de teste. A Figura 23 sumariza as estimativas de desempenho para cada algoritmo nas principais métricas utilizadas.

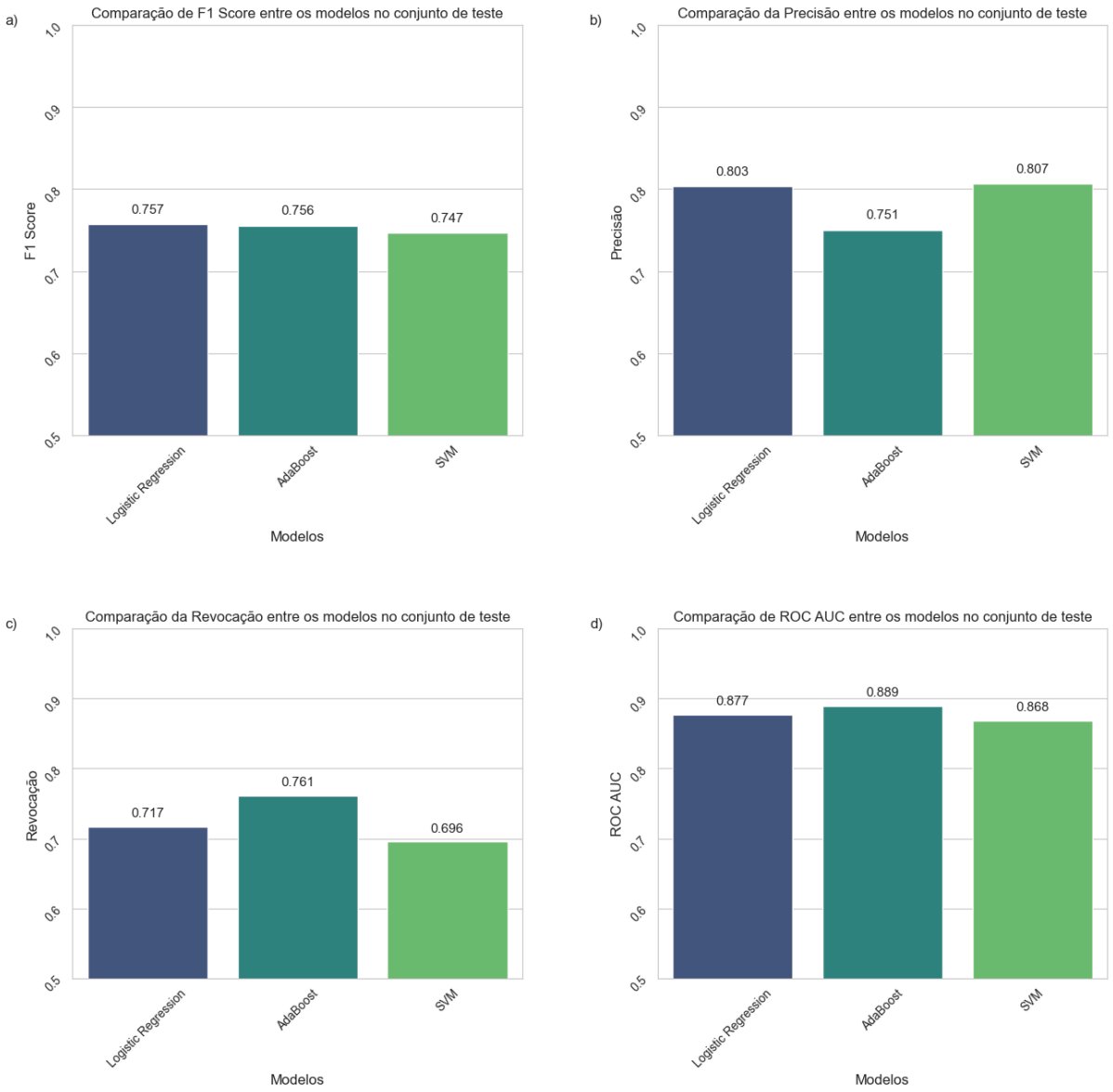


Figura 23 - Comparação da estimativa de desempenho dos algoritmos com seus melhores hiperparâmetros. a) *F1-Score* de cada modelo no conjunto de teste. b) Precisão de cada modelo no conjunto de teste. c) Revocação de cada modelo no conjunto de teste. d) Área sob a curva ROC de cada modelo no conjunto de teste.

A Figura 23a apresenta uma diferença de aproximadamente 0,009 entre o maior e o menor *F1-Score*, sendo que a diferença da métrica para a Regressão Logística e AdaBoost é de aproximadamente 0,001.

Em termos de precisão e revocação, sumarizados na Figura 23b e Figura 23c, o desempenho segue a mesma tendência apresentada pela distribuição na otimização de hiperparâmetros, onde o AdaBoost tem a menor precisão mas a maior revocação, onde todos os modelos compensam a relação de compromisso entre as duas métricas.

A área sob a curva ROC, apresentada na Figura 23c, pode ser utilizada como um critério de desempate na seleção do melhor modelo, pois é uma métrica importante para definir um bom classificador. Esta métrica varia 0,021 entre os algoritmos, sendo AdaBoost o modelo com maior pontuação.

A Figura 24 coloca o desempenho de cada modelo no conjunto de teste como um ponto na distribuição de cada métrica na otimização de hiperparâmetros, com objetivo de comparar a relação entre as tendências e sua proximidade da média.

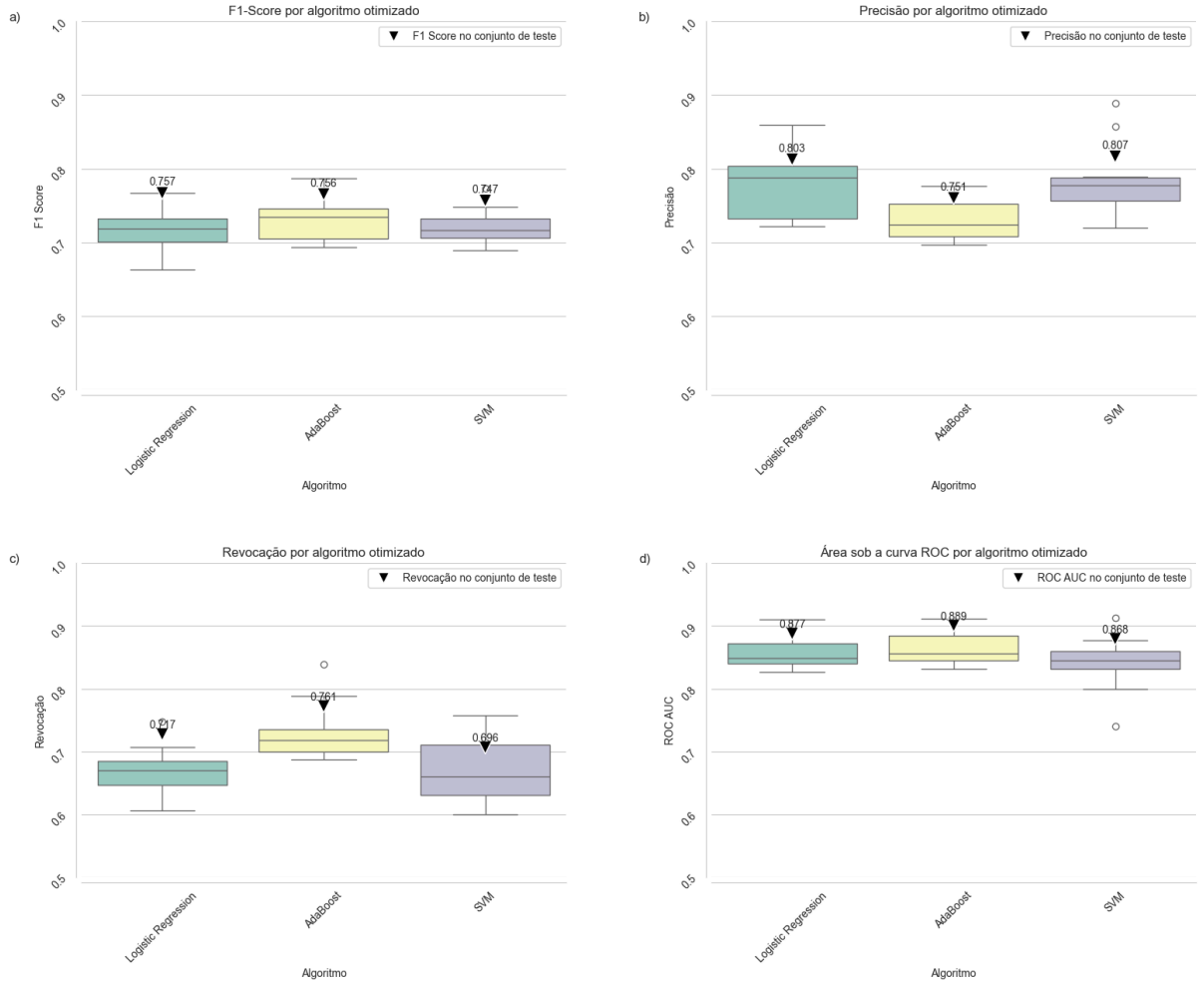


Figura 24 - Distribuição das métricas de avaliação na validação cruzada aninhada, indicando o ponto do desempenho de cada modelo avaliado nos dados de teste. a) Box-plot do F1-Score por algoritmo otimizado. b) Box-plot da precisão por algoritmo otimizado. c) Box-plot da revocação por algoritmo otimizado. d) Box-plot da área sob a curva ROC por algoritmo otimizado.

De forma individual, cada modelo apresentou um perfil de curva ROC muito similar, assim como o valor da área sob esta curva. A Figura 25 apresenta a curva ROC para cada modelo avaliado no conjunto de teste.

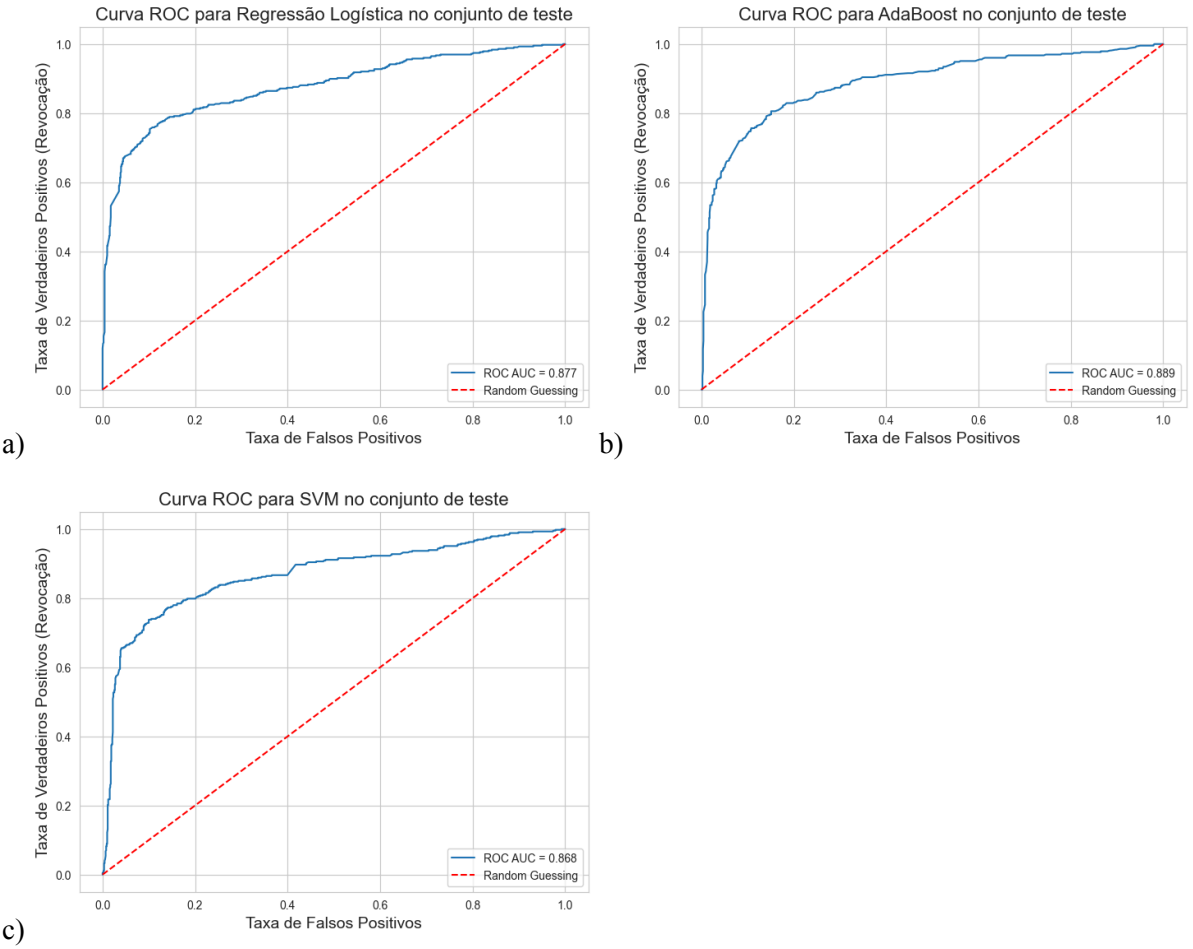


Figura 25 - Curva ROC para cada modelo avaliado no conjunto de teste. a) Regressão Logística. b) AdaBoost. c) SVM.

A Figura 26 mostra a matriz de confusão de cada modelo avaliado no conjunto de teste. O modelo de Regressão Logística apresenta uma maior taxa de predições Falso-Negativas (FN) do que Falso-Positivas (FP), sendo pouco mais de 9% FN e mais de 5% FP, como mostra a Figura 26a. Já o AdaBoost apresenta uma taxa similar de predições FP e FN sendo aproximadamente 8% das predições cada, conforme indica a Figura 26b. Por fim, o modelo SVM também apresenta uma maior taxa de predições FN, sendo quase 10% das predições, em relação à taxa de predições FP que representam pouco mais de 5% das predições. Esta relação é refletida na métrica de *F1-Score*, observada na Figura 23a, que pondera a relação entre precisão e revocação para conjuntos de dados desbalanceados.

Nos três modelos, a classe negativa é a classe mais prevista corretamente, pois é a classe que corresponde à permanência, classe majoritária no conjunto de dados desbalanceado.

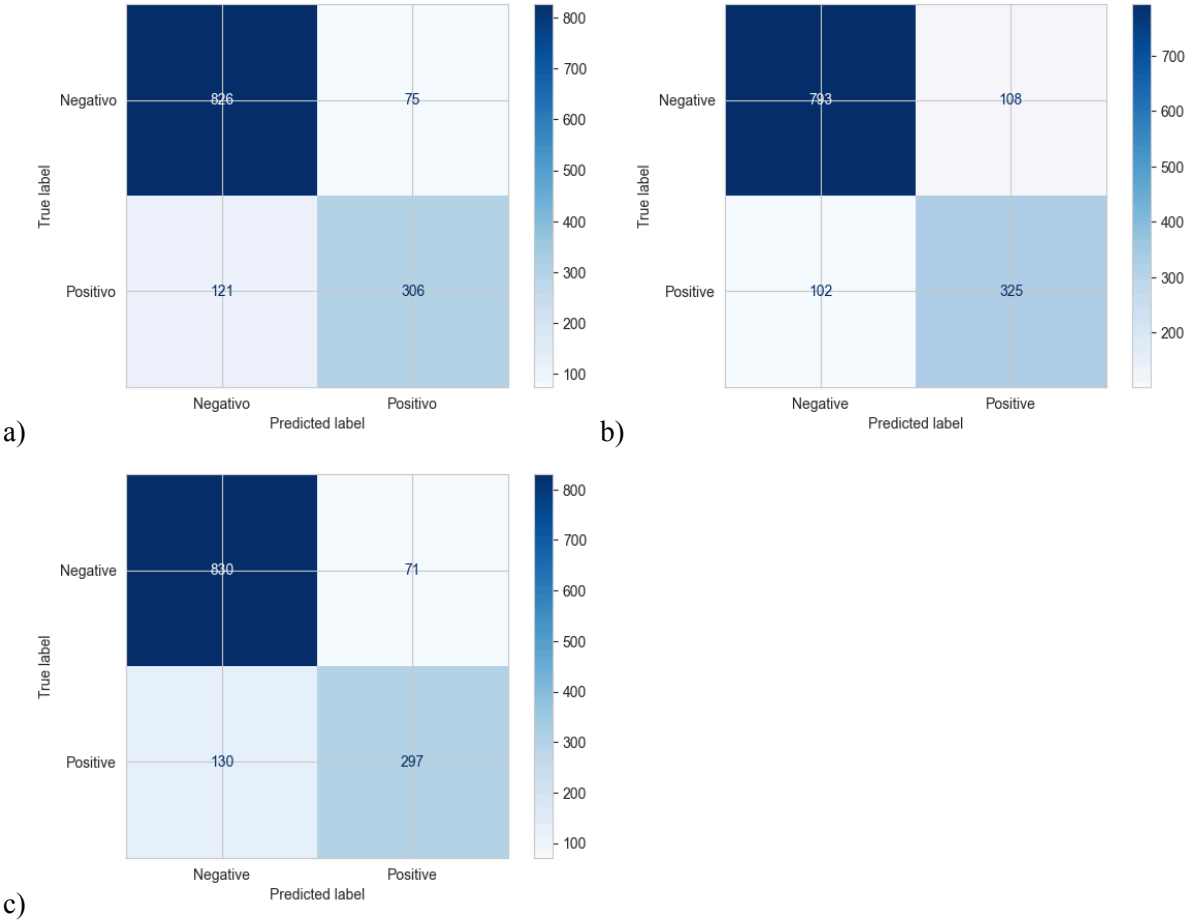


Figura 26 - Matrizes de confusão para cada modelo avaliado no conjunto de teste. a) Regressão Logística. b) AdaBoost. c) SVM.

Para selecionar o melhor modelo, visto o ‘empate técnico’ entre os modelos, principalmente entre Regressão Logística e AdaBoost, utilizamos a métrica de área sob a curva ROC para a seleção final. Assim, o modelo AdaBoost foi o modelo selecionado para o treinamento final com todos os dados e análise de interpretabilidade.

7. Interpretação do melhor modelo treinado

Para interpretar e explicar as decisões tomadas, treinamos o modelo selecionado AdaBoost em todos os dados do conjunto. Visto que se trata de um algoritmo *ensemble*, baseado em *boosting*, não são modelos que apresentam uma interpretabilidade intrínseca. Ou seja, não carregam em sua estrutura e parâmetros a capacidade de serem explicados ou apresentados em termos compreensíveis para seres humanos. Portanto, utilizamos técnicas de interpretabilidade extrínseca, que possibilitam a explicabilidade de um modelo através de uma coleção de artefatos visuais e/ou interativos que fornecem ao usuário uma descrição suficiente do comportamento de um modelo. [17]

Dentre estas técnicas e análises *post-hoc*, foi realizada a análise de interpretação para o modelo AdaBoost com Importância de Atributos, *Partial Dependence Plots (PDP)* e *SHapley Additive exPlanations (SHAP)*.

7.1 Importância de Atributos

Utilizou-se um método de Importância de Atributos com testes de permutação, pois é uma abordagem agnóstica a modelos. Este método mede o aumento no erro de predição do modelo ao permutarmos os valores do atributo, distorcendo a relação real entre cada atributo e a saída esperada. Um atributo é importante se, ao embaralhar seus valores, aumentamos o erro do modelo, pois indica que o modelo usa este atributo para realizar a predição [17]. A Figura 27 apresenta o gráfico da Importância de Atributos para o modelo AdaBoost.

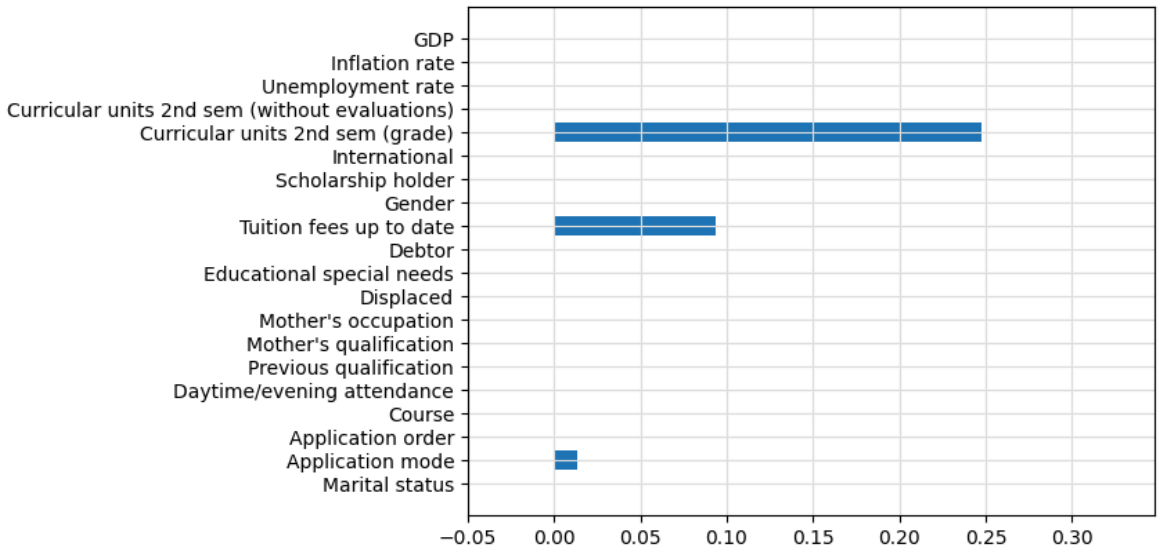


Figura 27 - Importância de Atributos com testes de permutação para o modelo AdaBoost.

O atributo *Curricular units 2nd sem (grade)* destaca-se como o atributo com maior importância, apresentando uma pontuação de aproximadamente 0,25. Em seguida, o atributo *Tuition fees up to date* apresenta aproximadamente 0,09 e *Application mode* apresenta pouco mais de de 0,01 na pontuação. É importante lembrar que a *pipeline* de modelagem inclui um passo de seleção de atributos, onde aparecem todos estes atributos, assim como *Scholarship holder*. Para o AdaBoost, este último atributo selecionado apresentou uma importância de 0,00 na pontuação estimada.

7.2 Partial Dependence Plot (PDP)

Esta técnica analisa o efeito marginal que um ou dois atributos têm sobre o resultado previsto de um modelo, podendo mostrar graficamente a relação entre a saída e cada atributo. Para um determinado valor do atributo, representa a previsão média do modelo para o caso em que forçamos todas as instâncias a assumirem aquele valor, permitindo se analisar uma relação causal entre atributo e saída no modelo [17].

As Figuras 28a e 28b apresentam os *Partial Dependence Plots* para os atributos do modelo AdaBoost.

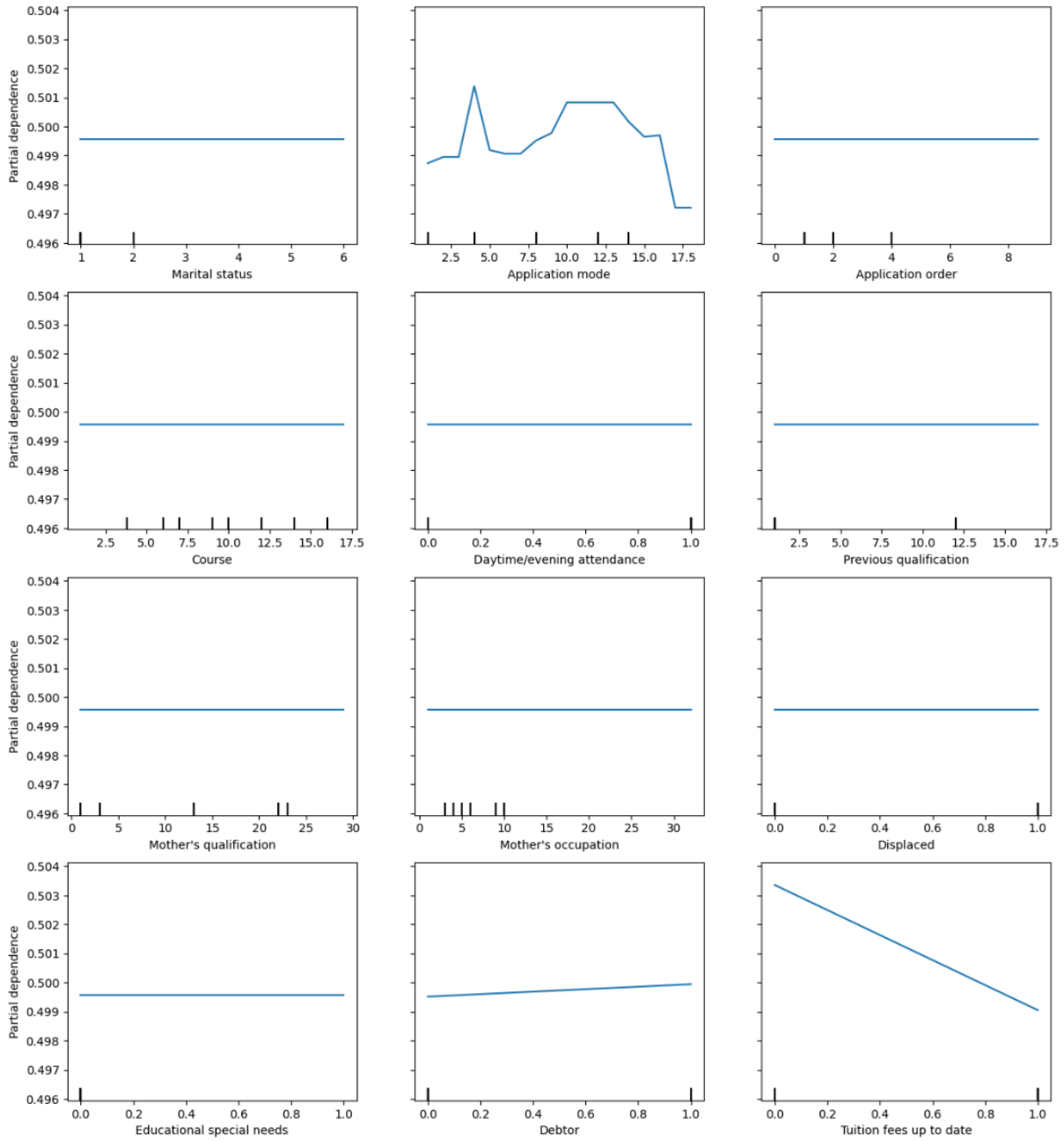


Figura 28a - *Partial Dependence Plot* para os atributos do modelo AdaBoost.

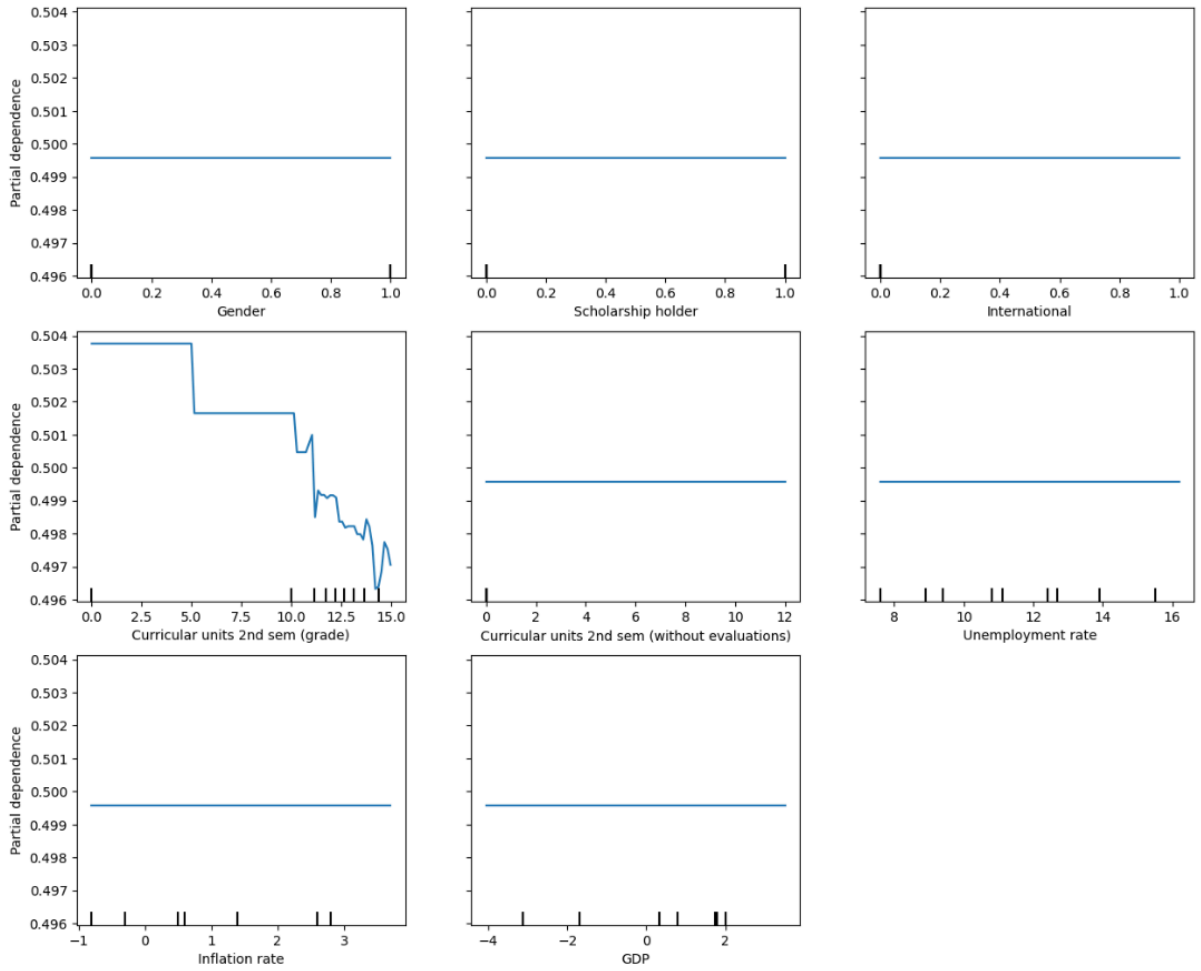


Figura 28b - *Partial Dependence Plot* para os atributos do modelo AdaBoost.

Observa-se que, apesar de apresentar variações, o atributo *Curricular units 2nd sem (grade)* exibe um padrão que pode ser aproximado a uma relação linear, onde quanto maior a nota, menor a dependência desta variável para predizer a instância como evasão. Já o atributo *Tuition fees up to date* apresenta uma relação linear, pois é um atributo binário, onde a instância de um aluno está com a mensalidade (*Tuition*) em dia, menor a dependência desta variável para a predição como evasão. O atributo *Application mode* apresenta um padrão de dependência não-linear, porém como é um atributo categórico codificado em diversos sub-atributos, é difícil interpretá-lo. Trabalhos futuros podem transformar este atributo categórico codificado em vários atributos categóricos binários, permitindo sua interpretação individualizada.

7.3 SHapley Additive exPlanations (SHAP)

O método SHAP tem como objetivo explicar a previsão de uma instância calculando um valor de Shapley para a contribuição de cada atributo para a previsão. Este valor, baseado na teoria de jogos de coalizão, é a contribuição marginal média de um valor de atributo em todas as coalizões possíveis [18]. Utilizou-se o *Kernel SHAP* como implementação viável e agnóstica a modelos para a explicação de um modelo AdaBoost, pois a biblioteca *shap* não disponibiliza um método específico para a explicação deste tipo de algoritmo preditivo. A Figura 29 apresenta gráfico de valores de SHAP para cada atributo selecionado pela *pipeline* do modelo AdaBoost.

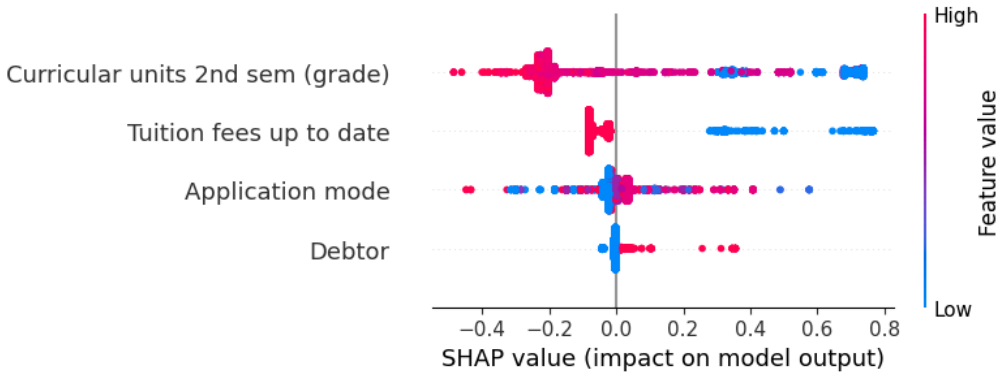


Figura 29 - Gráfico com valores de SHAP para cada atributo.

Para o atributo *Curricular units 2nd sem (grade)*, quanto maior o valor das notas, tende a apresentar um impacto negativo na predição do modelo (0.0, -0.5), ou seja, tende à predição de permanência. De forma contrária, mas com menos pontos por conta do desbalanceamento, quanto menor o valor das notas, tende a apresentar um impacto positivo na predição do modelo, tendendo à classificação como evasão. Porém, existe uma intersecção significativa de pontos onde, uma nota tendendo a aumentar pode também ter impacto positivo na predição de evasão.

Para *Tuition feeds up to date*, com valores de SHAP entre 0.0 e -0.2, instâncias com valor igual a 1 (pagamento em dia) apresentam impacto negativo na predição da evasão. Já no intervalo de valores SHAP entre 0.2 e 0.8, instâncias com valor igual a 0 (pagamento atrasado) apresentam impacto positivo na predição da evasão. Este atributo tem uma separação bem delimitada em termos de valor SHAP e seus valores binários, indicando que podem separar bem os dados.

Para *Application mode* existem muitas intersecções entre valores, pois codificam outras categorias binárias, dificultando sua explicação.

Para *Debtor*, no intervalo de valores SHAP entre 0.0 e -0.1, instâncias com valor igual a 0 (não devedor, a partir de dados socioeconômicos) apresentam impacto negativo na predição de evasão. Porém, este intervalo e impacto negativo é menor do que o atributo *Tuition fees up to date*. No intervalo de valores SHAP entre 0.0 e 0.4, instâncias com valor igual a 1 (devedores, a partir de dados socioeconômicos) apresentam impacto positivo na predição de evasão. Da mesma forma, tem um impacto menor do que *Tuition fees up to date*. Porém, de forma similar, também tem uma separação bem delimitada de valor SHAP para seus valores binários.

8. Considerações finais

Neste trabalho, exploramos os dados e diversos modelos preditivos para predição da evasão de estudantes universitários. Através da primeira etapa, composta por análise exploratória, pré-processamento dos dados e *spot-checking* de algoritmos, preparamos um conjunto de dados adequado e selecionamos três modelos potenciais para a modelagem da nossa tarefa preditiva. Em uma segunda etapa, realizamos a otimização de hiperparâmetros, através de validação cruzada aninhada. Ao validar no conjunto de teste, verificamos que os três algoritmos apresentam um desempenho semelhante entre si, o que na interpretação extrínseca dos atributos pode ser explicado através do número baixo de atributos com contribuição significativa na predição. Assim, selecionamos e treinamos um modelo AdaBoost, através de uma *pipeline de modelagem* composta por seleção de atributos, balanceamento de dados e algoritmo preditivo, utilizando os melhores hiperparâmetros conforme avaliação através das métricas selecionadas.

O desempenho principal deste modelo de aproximadamente 75% em *F1-score* mostra uma capacidade aceitável de se utilizar este modelo na predição de novos dados, pois também apresenta um bom nível de área sob a curva ROC, de aproximadamente 89%, o que é uma métrica importante para um modelo preditivo de classificação.

Como possibilidade futura, algoritmos com outros vieses indutivos, como Redes Neurais ou *XGBoost*, podem ser adicionados ao *spot-checking*, permitindo investigar um leque maior de possibilidades. Todavia, um dos pontos de melhoria mais importantes se refere aos dados. O atributo *Application mode* deve ser transformado, através de técnicas como *One-hot encoding*, para diversos atributos categóricos binários. Neste trabalho, este atributo foi utilizado de forma codificada, apresentando 18 categorias diferentes. Na análise de importância e interpretabilidade, apareceu com boa pontuação, mas este problema não permitiu sua explicabilidade. Por fim, outra possibilidade relativa aos dados seria a adição de mais dados acadêmicos, como notas de outros semestres além dos dois primeiros, assim como a intersecção entre dados socioeconômicos e acadêmicos e dados relativos à saúde, visto o impacto da saúde mental na performance acadêmica.

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] GEEKSFORGEEKS. 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] GEEKSFORGEEKS. 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] GEEKSFORGEEKS. 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.

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	$\mathbb{N}$
Dados socioeconômicos			
Atributo	**Domínio**	**Descrição**	**Categoria(s)**
Mother's qualification / Father's qualification	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 3—Higher Education—degree 4—Higher Education—master's degree 5—Higher Education—doctorate 6—Frequency of Higher Education 7—12th Year of Schooling—not completed 8—11th Year of Schooling—not completed 9—7th Year (Old) 10—Other—11th Year of Schooling

			11—2nd year complementary high school course 12—10th Year of Schooling 13—General commerce course 14—Basic Education 3rd Cycle (9th/10th/11th Year) or Equivalent 15—Complementary High School Course 16—Technical-professional course 17—Complementary High School Course—not concluded 18—7th year of schooling 19—2nd cycle of the general high school course 20—9th Year of Schooling—not completed 21—8th year of schooling 22—General Course of Administration and Commerce 23—Supplementary Accounting and Administration 24—Unknown 25—Cannot read or write 26—Can read without having a 4th year of schooling 27—Basic education 1st cycle (4th/5th year) or equivalent 28—Basic Education 2nd Cycle (6th/7th/8th Year) or equivalent 29—Technological specialization course 30—Higher education—degree (1st cycle) 31—Specialized higher studies course 32—Professional higher technical course 33—Higher Education—master's degree (2nd cycle) 34—Higher Education—doctorate (3rd cycle)
Mother's occupation / Father's occupation	Categórico nominal (codificado)	A ocupação profissional pais do estudante	1—Student 2—Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers 3—Specialists in Intellectual and Scientific Activities 4—Intermediate Level Technicians and Professions 5—Administrative staff 6—Personal Services, Security and Safety Workers, and Sellers 7—Farmers and Skilled Workers in Agriculture, Fisheries, and Forestry 8—Skilled Workers in Industry, Construction, and Craftsmen 9—Installation and Machine Operators and Assembly Workers 10—Unskilled Workers 11—Armed Forces Professions 12—Other Situation; 13—(blank) 14—Armed Forces Officers 15—Armed Forces Sergeants 16—Other Armed Forces personnel 17—Directors of administrative and commercial services 18—Hotel, catering, trade, and other services directors 19—Specialists in the physical sciences, mathematics, engineering, and related techniques 20—Health professionals 21—Teachers

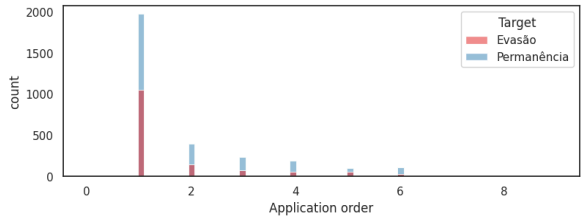
			22—Specialists in finance, accounting, administrative organization, and public and commercial relations 23—Intermediate level science and engineering technicians and professions 24—Technicians and professionals of intermediate level of health 25—Intermediate level technicians from legal, social, sports, cultural, and similar services 26—Information and communication technology technicians 27—Office workers, secretaries in general, and data processing operators 28—Data, accounting, statistical, financial services, and registry-related operators 29—Other administrative support staff 30—Personal service workers 31—Sellers 32—Personal care workers and the like 33—Protection and security services personnel 34—Market-oriented farmers and skilled agricultural and animal production workers 35—Farmers, livestock keepers, fishermen, hunters and gatherers, and subsistence 36—Skilled construction workers and the like, except electricians 37—Skilled workers in metallurgy, metalworking, and similar 38—Skilled workers in electricity and electronics 39—Workers in food processing, woodworking, and clothing and other industries and crafts 40—Fixed plant and machine operators 41—Assembly workers 42—Vehicle drivers and mobile equipment operators 43—Unskilled workers in agriculture, animal production, and fisheries and forestry 44—Unskilled workers in extractive industry, construction, manufacturing, and transport 45—Meal preparation assistants 46—Street vendors (except food) and street service providers
Displaced	Categórico binário (codificado)	Indica se uma pessoa foi forçada a abandonar sua casa ou local de nascimento devido a conflitos, violência, violação de direitos humanos, etc.	0: Não 1: Sim
Educational special needs	Categórico binário (codificado)	Se o estudante possui alguma necessidade educacional especial	0: Não 1: Sim

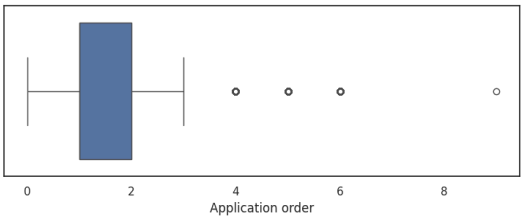
International	Categórico binário (codificado)	Indica se o estudante é de outro país	0: Não 1: Sim
Debtor	Categórico binário (codificado)	Indica se a pessoa estudante é devedora ou não.	0: Não 1: Sim
Tuition fees up to date	Categórico binário (codificado)	Pagamento da universidade em dia	0: Não 1: Sim
Scholarship holder	Categórico binário (codificado)	Bolsista	0: Não 1: Sim
Dados acadêmicos			
Application mode	Categórico nominal (codificado)	Modo de aplicação para entrada na universidade.	1: 1st phase—general contingent 2: Ordinance No. 612/93 3: 1st phase—special contingent (Azores Island) 4: Holders of other higher courses 5: Ordinance No. 854-B/99 6: International student (bachelor) 7: 1st phase—special contingent (Madeira Island) 8: 2nd phase—general contingent 9: 3rd phase—general contingent 10: Ordinance No. 533-A/99, item b2) (Different Plan) 11: Ordinance No. 533-A/99, item b3 (Other Institution) 12: Over 23 years old 13: Transfer 14: Change in course 15: Technological specialization diploma holders 16: Change in institution/course 17: Short cycle diploma holders 18: Change in institution/course (International)
Application order	Numérico ordinal	Ordem de prioridade numérica da aplicação do estudante.	$\mathbb{N}$
Daytime/evening attendance	Categórico binário (codificado)	Indica se o estudante frequenta aulas durante o dia ou à noite.	1: daytime 0: evening
Previous qualification	Categórico nominal (codificado)	Qualificações obtidas antes do ingresso no ensino superior.	1—Secondary education 2—Higher education—bachelor’s degree 3—Higher education—degree 4—Higher education—master’s degree 5—Higher education—doctorate 6—Frequency of higher education 7—12th year of schooling—not completed

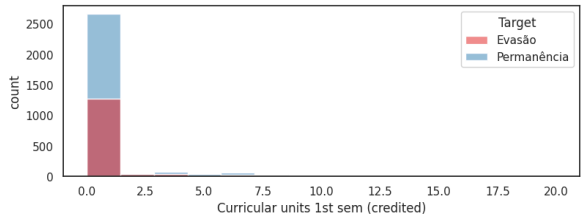
			8—11th year of schooling—not completed 9—Other—11th year of schooling 10—10th year of schooling 11—10th year of schooling—not completed 12—Basic education 3rd cycle (9th/10th/11th year) or equivalent 13—Basic education 2nd cycle (6th/7th/8th year) or equivalent 14—Technological specialization course 15—Higher education—degree (1st cycle) 16—Professional higher technical course 17—Higher education—master's degree (2nd cycle)
Course	Categórico nominal (codificado)	Curso escolhido pelo estudante.	1: Biofuel Production Technologies 2: Animation and Multimedia Design 3: Social Service (evening attendance) 4: Agronomy 5: Communication Design 6: Veterinary Nursing 7: Informatics Engineering 8: Equiniculture 9: Management 10: Social Service 11: Tourism 12: Nursing 13: Oral Hygiene 14: Advertising and Marketing Management 15: Journalism and Communication 16: Basic Education 17: Management (evening attendance)
Curricular units 1st sem (enrolled)	Numérico discreto	Número de disciplinas que o indivíduo se matriculou no primeiro semestre	$\mathbb{N}$
Curricular units 1st sem (approved)	Numérico discreto	Números de disciplinas que o estudante foi aprovado no primeiro semestre	$\mathbb{N}$
Curricular units 1st sem (grade)	Numérico contínuo	Nota do estudante ao final do primeiro semestre	$\mathbb{N}$
Curricular units 2nd sem (enrolled)	Numérico discreto	Número de disciplinas que o indivíduo se matriculou no segundo semestre	$\mathbb{N}$
Curricular units 2nd sem (approved)	Numérico discreto	Números de disciplinas que o estudante foi	$\mathbb{N}$

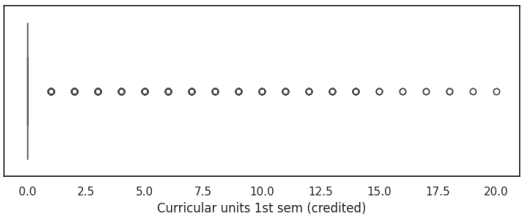
		aprovado no segundo semestre	
Curricular units 2nd sem (grade)	Numérico contínuo	Nota do estudante ao final do segundo semestre	$\mathbb{R}$
Fatores macroeconômicos			
Unemployment rate	Numérico contínuo	Taxa de desemprego do país de nacionalidade	$\mathbb{R}$
Inflation rate	Numérico contínuo	Taxa de inflação do país de nacionalidade	$\mathbb{R}$
GDP	Numérico contínuo	Produto Interno Bruto (PIB) do país de nacionalidade	$\mathbb{R}$
Atributo alvo			
Target	Categórico discreto	Indica a situação acadêmica do estudante	• Graduate • Dropout • Enrolled

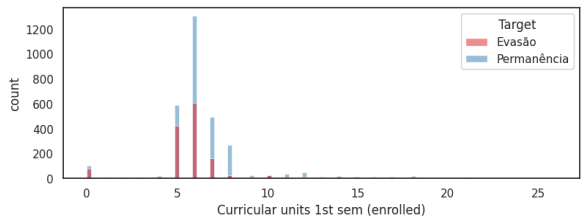
ANEXO 2 - DISTRIBUIÇÃO DE FREQUÊNCIA PARA ATRIBUTOS NUMÉRICOS

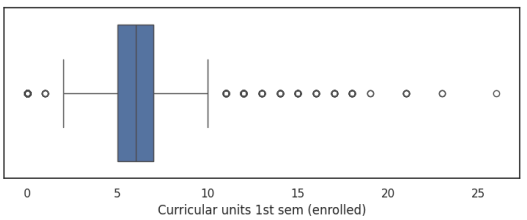


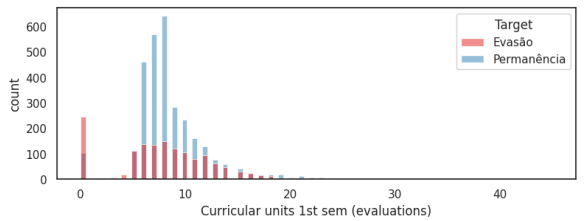


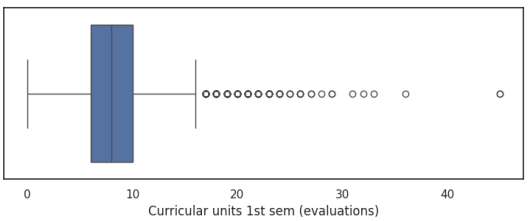


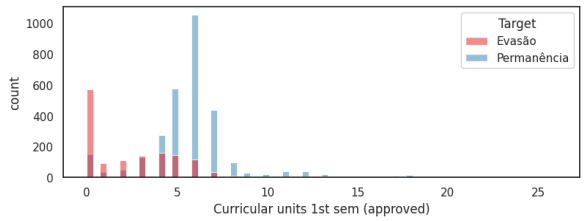


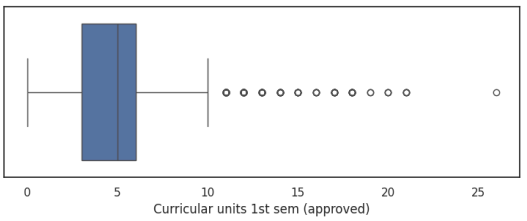


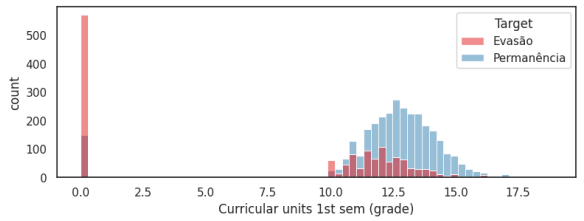


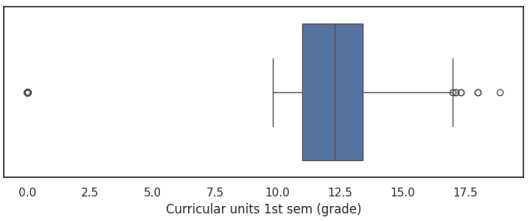


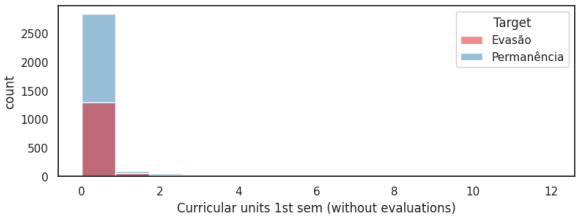


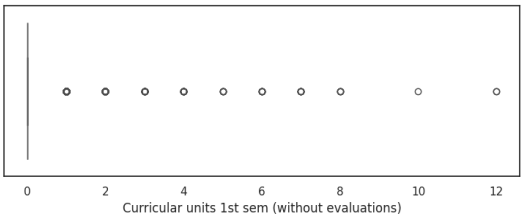


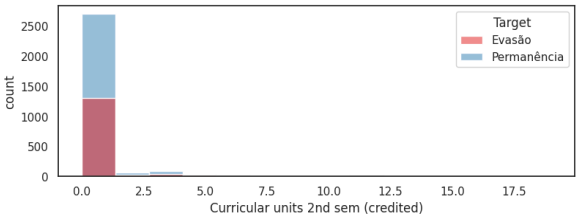


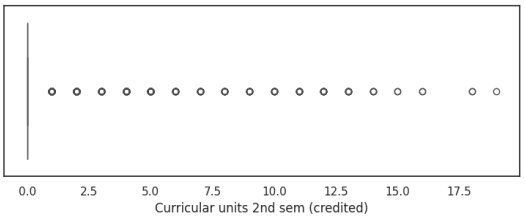


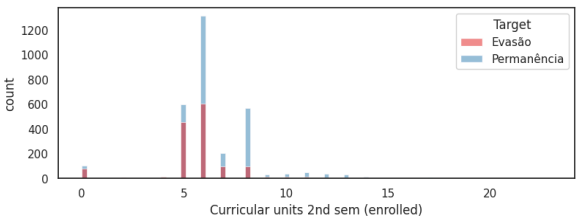


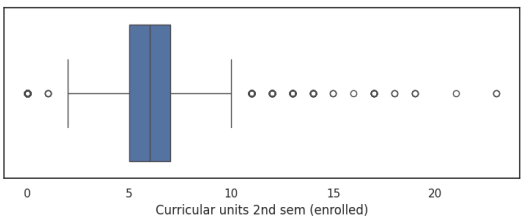

Target
Evasão
Permanência
count
Curricular units 1st sem (without evaluations)

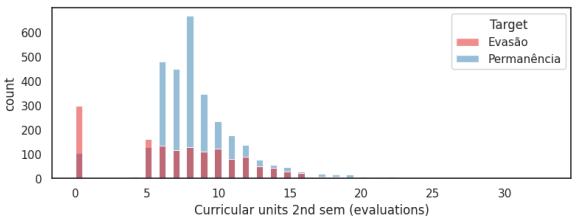

Curricular units 1st sem (without evaluations)

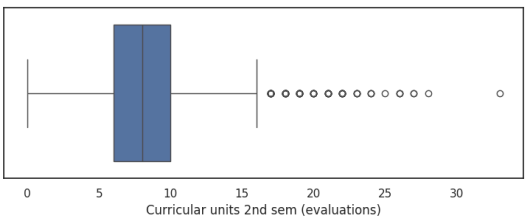

Target
Evasão
Permanência
count
Curricular units 2nd sem (credited)

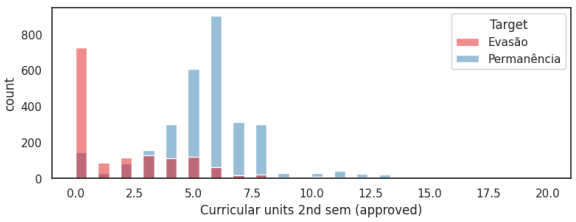

Curricular units 2nd sem (credited)

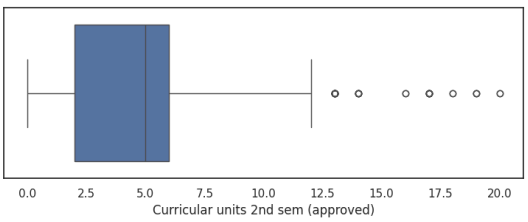

Target
Evasão
Permanência
count
Curricular units 2nd sem (enrolled)

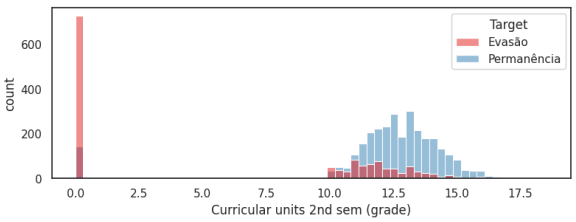

Curricular units 2nd sem (enrolled)

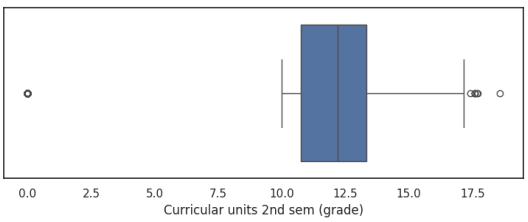

Target
Evasão
Permanência
count
Curricular units 2nd sem (evaluations)

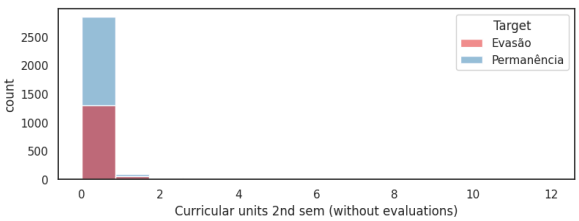

Curricular units 2nd sem (evaluations)

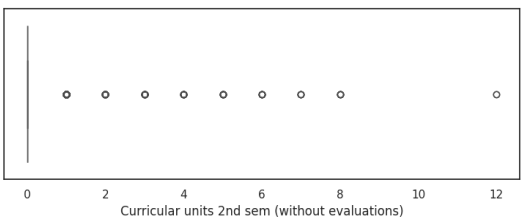

Target
Evasão
Permanência
count
Curricular units 2nd sem (approved)

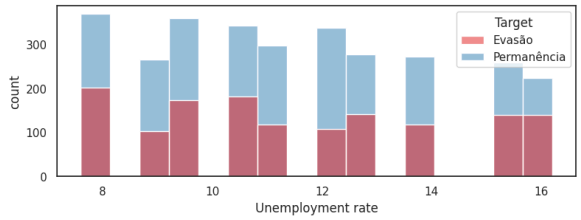

Curricular units 2nd sem (approved)

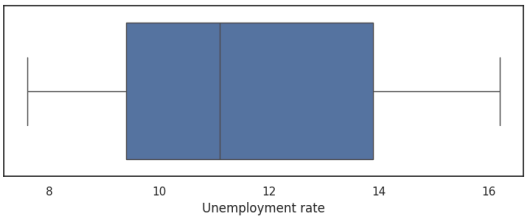

Target
Evasão
Permanência
count
Curricular units 2nd sem (grade)

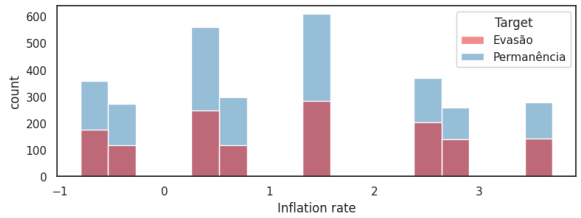

Curricular units 2nd sem (grade)

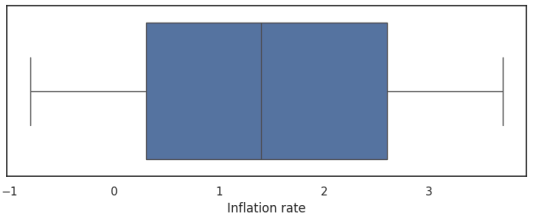

Target
Evasão
Permanência
count
Curricular units 2nd sem (without evaluations)

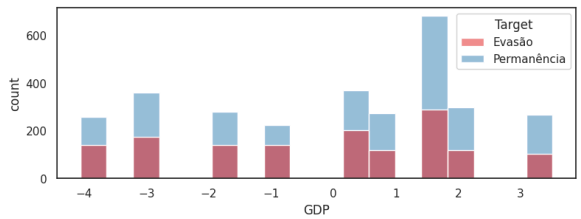

Curricular units 2nd sem (without evaluations)

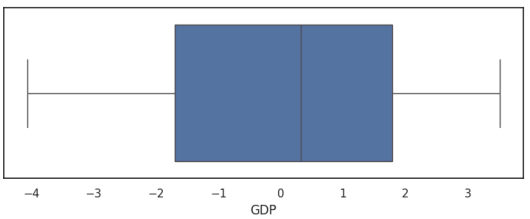

Target
Evasão
Permanência
count
Unemployment rate

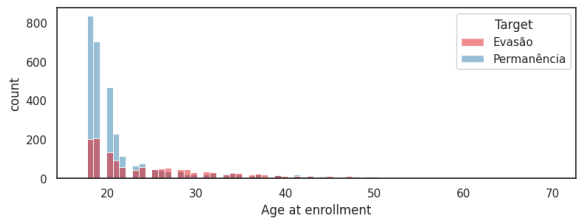

Unemployment rate

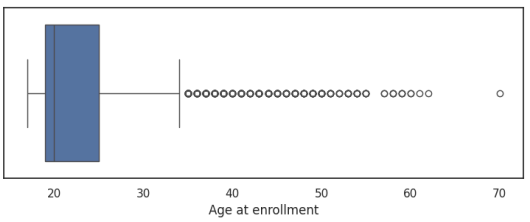

Target
Evasão
Permanência
count
Inflation rate


Inflation rate

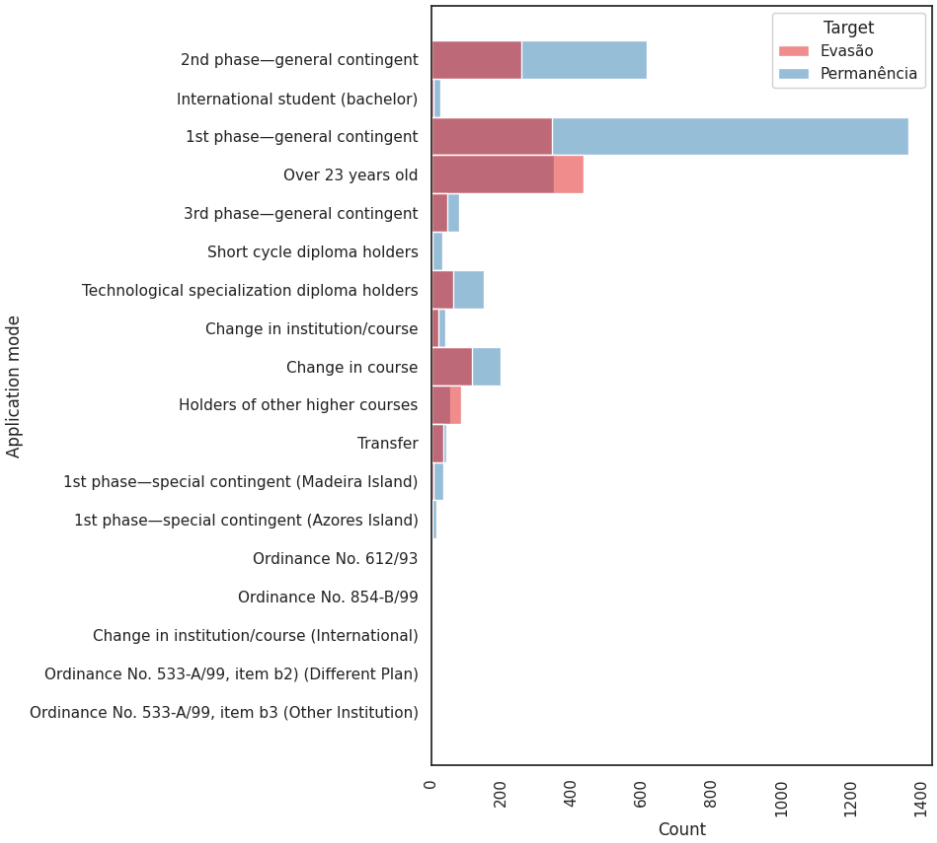

Target
Evasão
Permanência
count
GDP

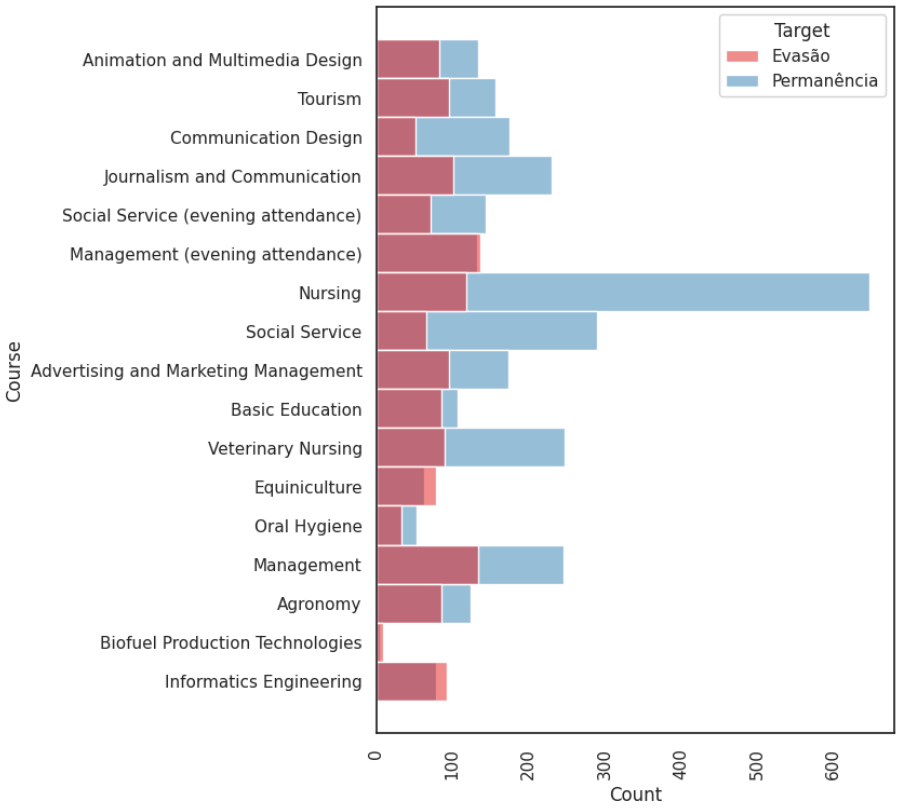

GDP

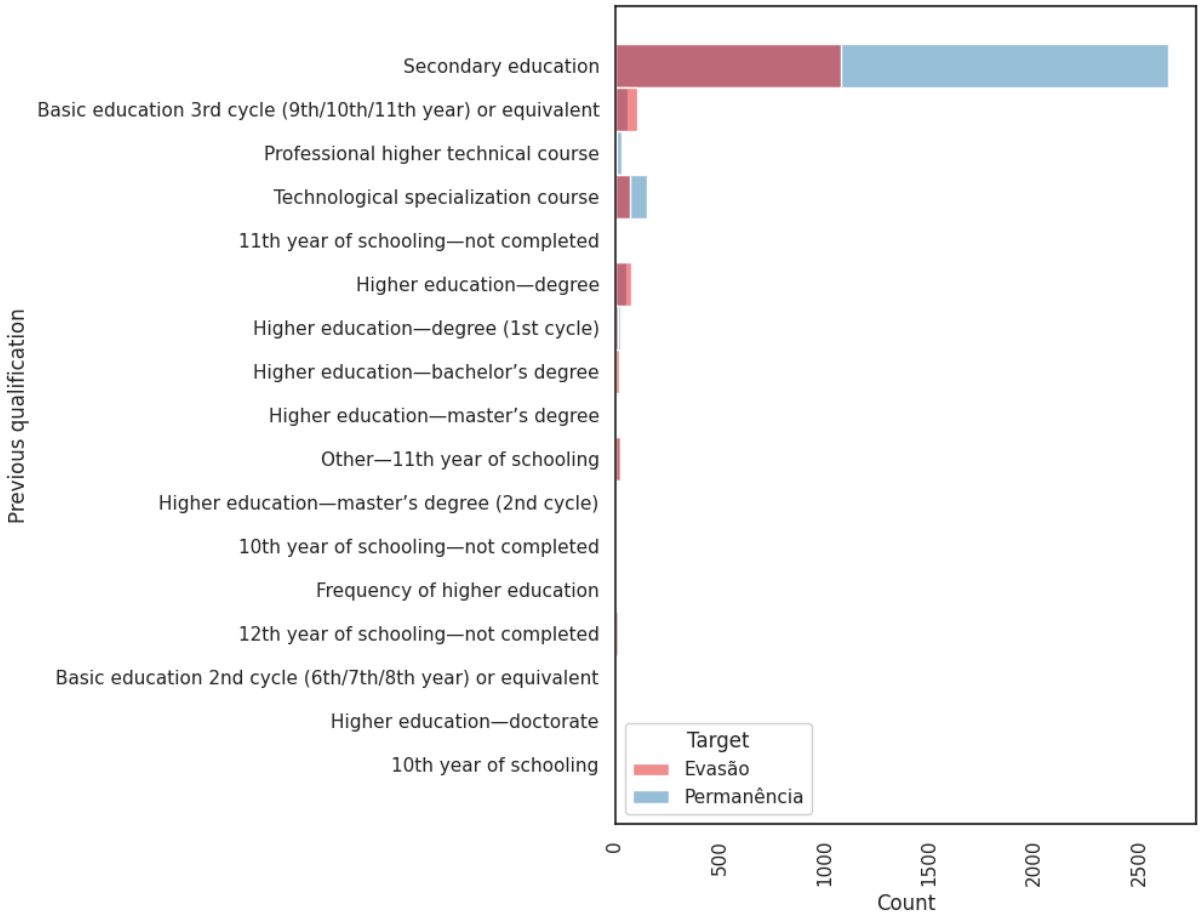

Target
Evasão
Permanência
count
Age at enrollment

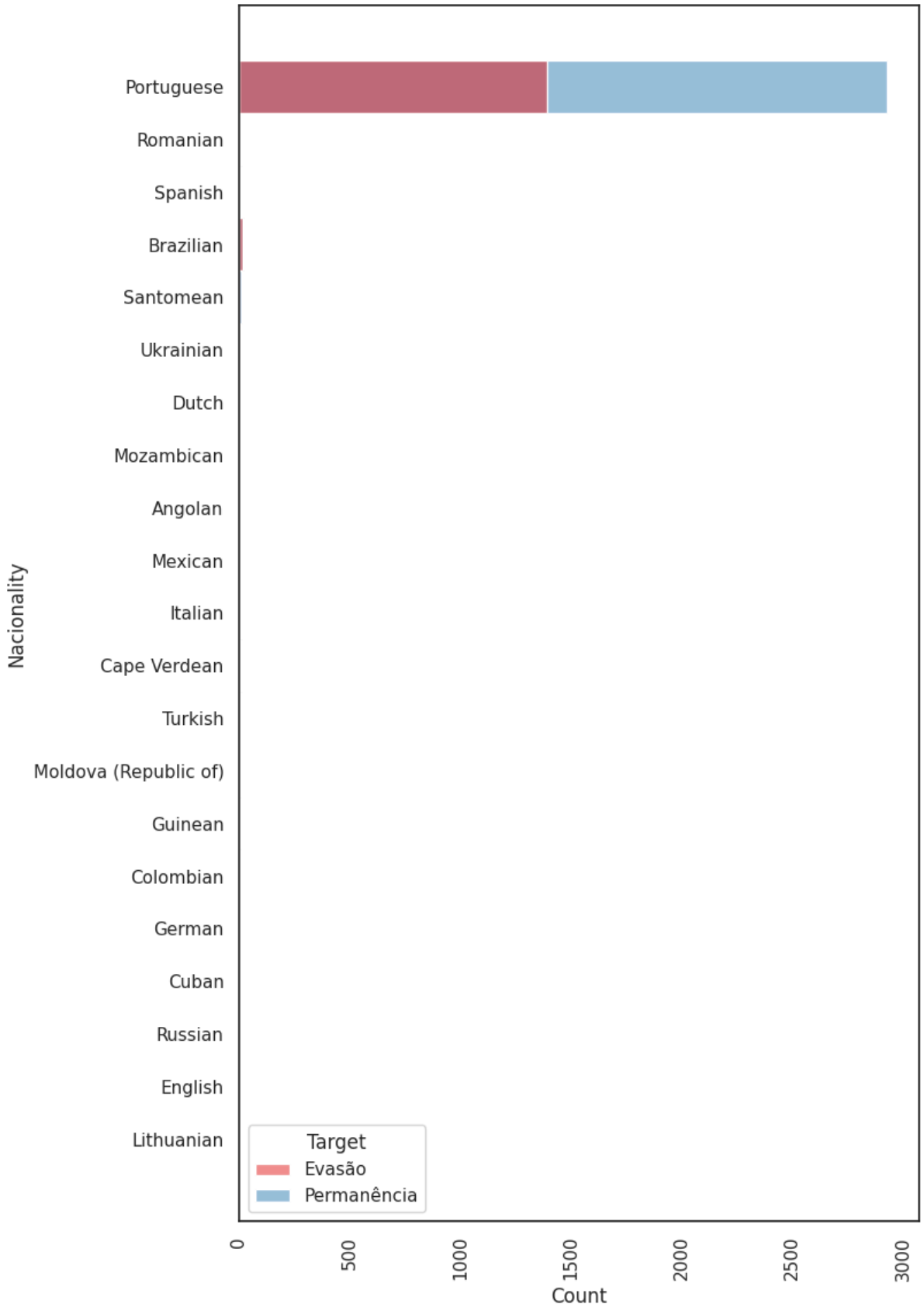

Age at enrollment

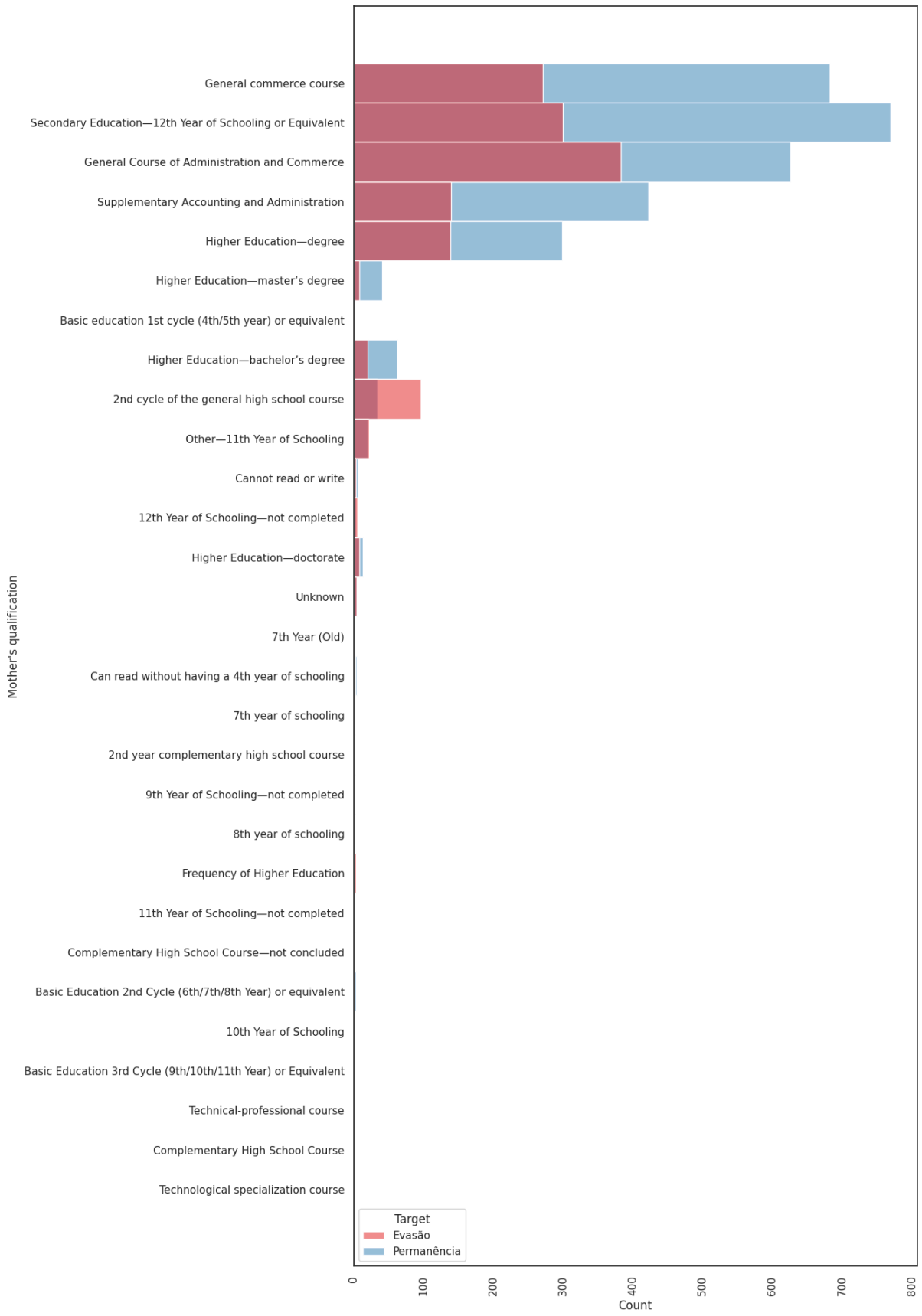
ANEXO 3 - DISTRIBUIÇÃO DE FREQUÊNCIA PARA ATRIBUTOS CATEGÓRICOS

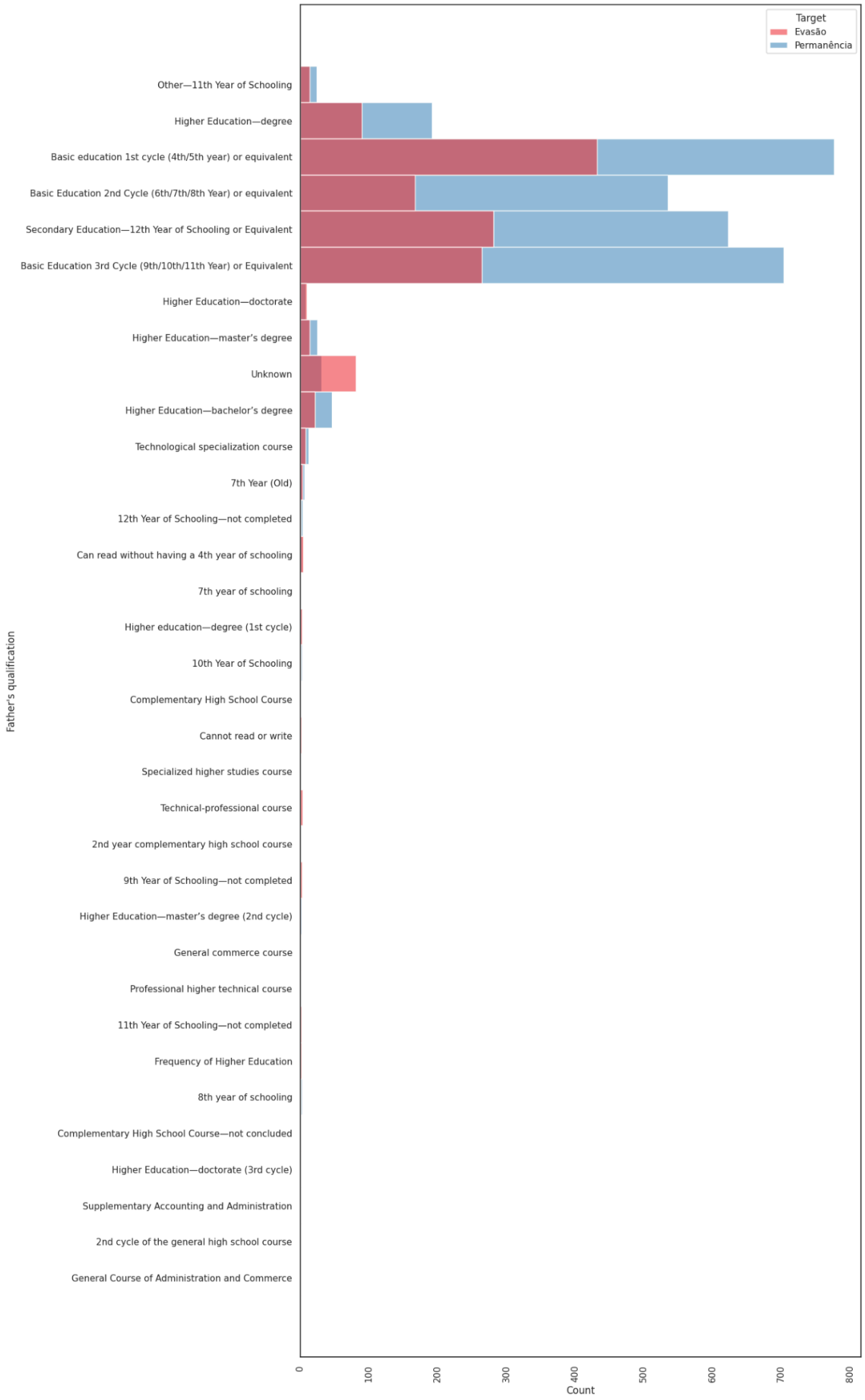


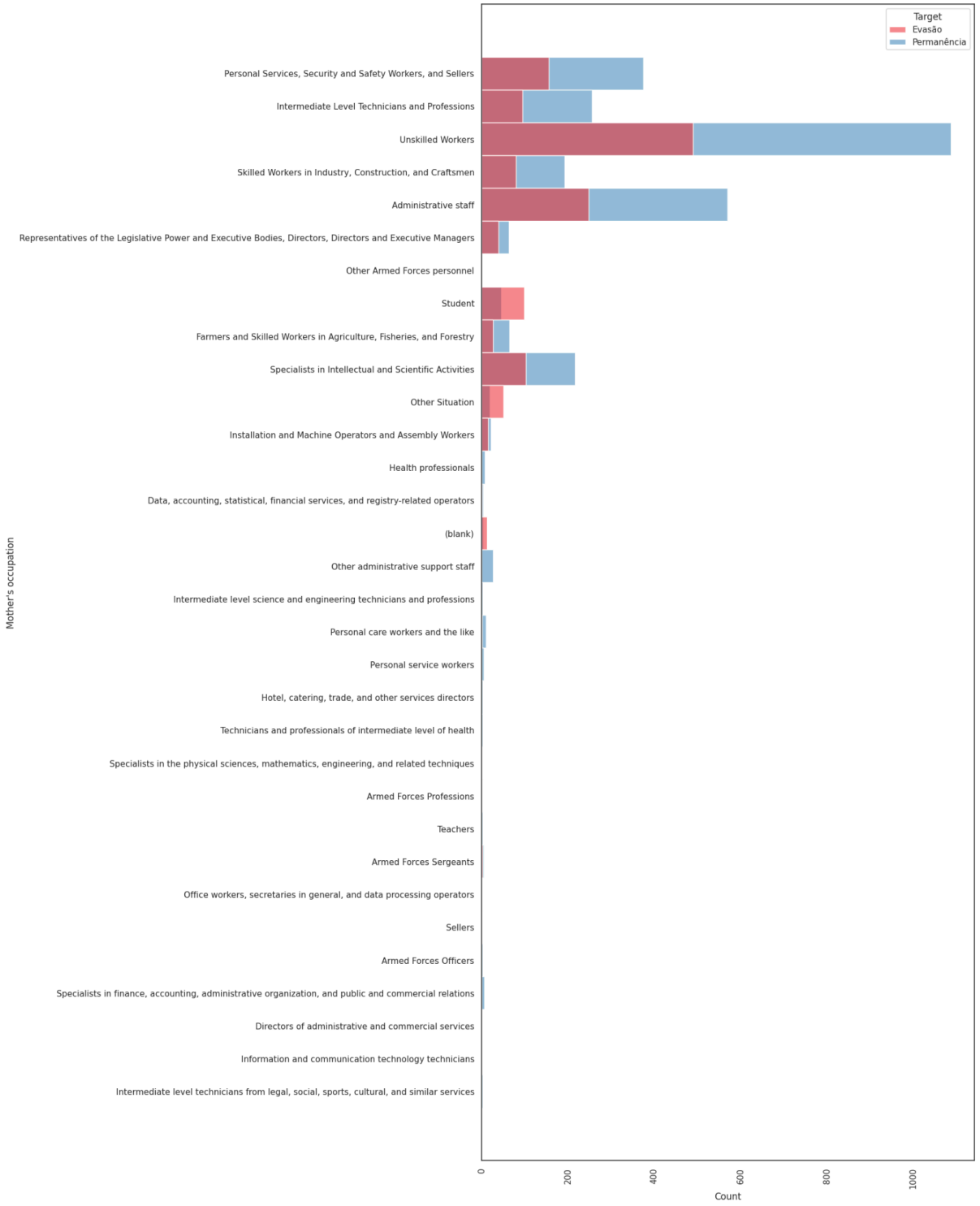


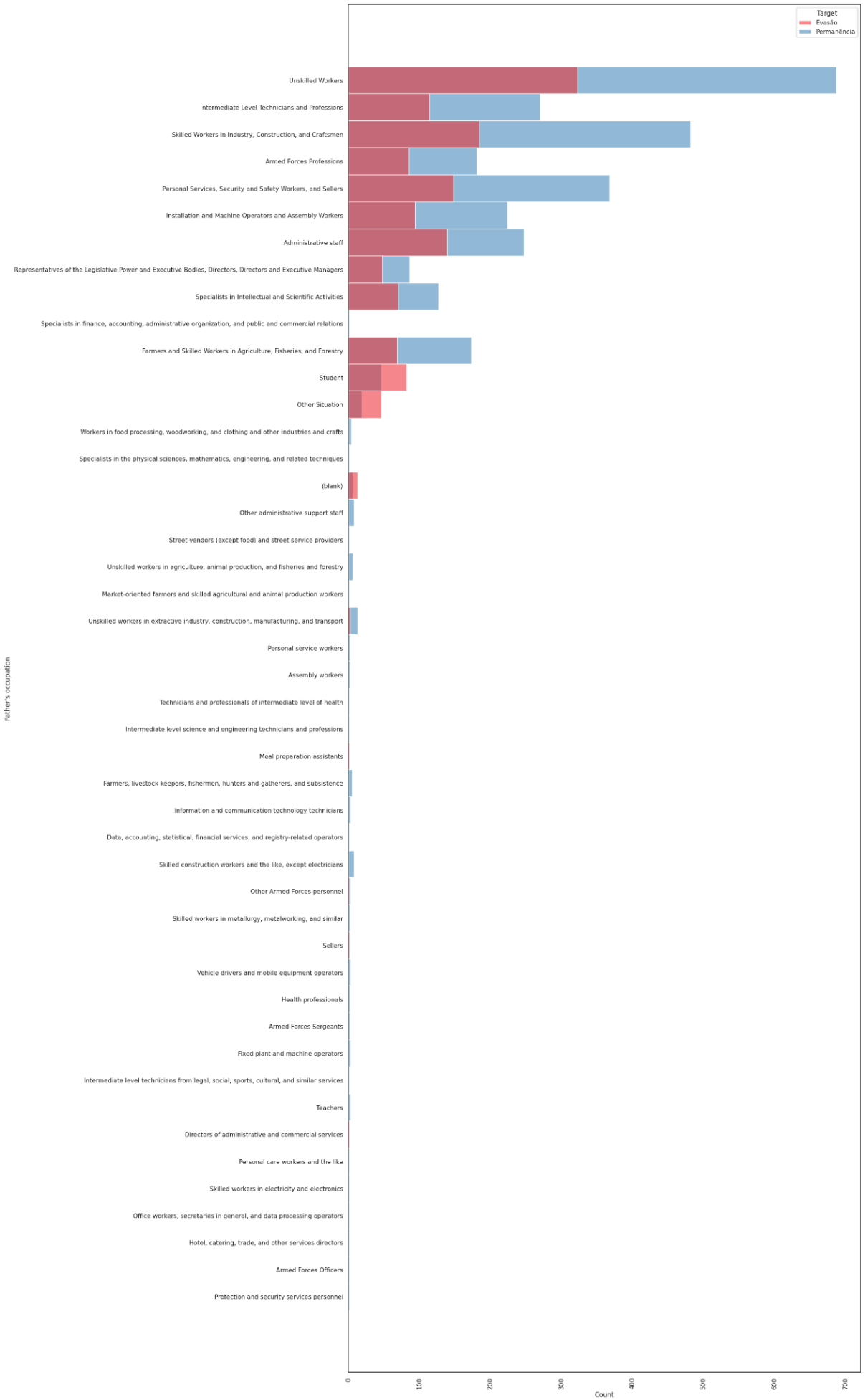

Previous qualification
Secondary education
Basic education 3rd cycle (9th/10th/11th year) or equivalent
Professional higher technical course
Technological specialization course
11th year of schooling—not completed
Higher education—degree
Higher education—degree (1st cycle)
Higher education—bachelor's degree
Higher education—master's degree
Other—11th year of schooling
Higher education—master's degree (2nd cycle)
10th year of schooling—not completed
Frequency of higher education
12th year of schooling—not completed
Basic education 2nd cycle (6th/7th/8th year) or equivalent
Higher education—doctorate
10th year of schooling
Target
Evasão
Permanência
0
500
1000
1500
2000
2500
Count

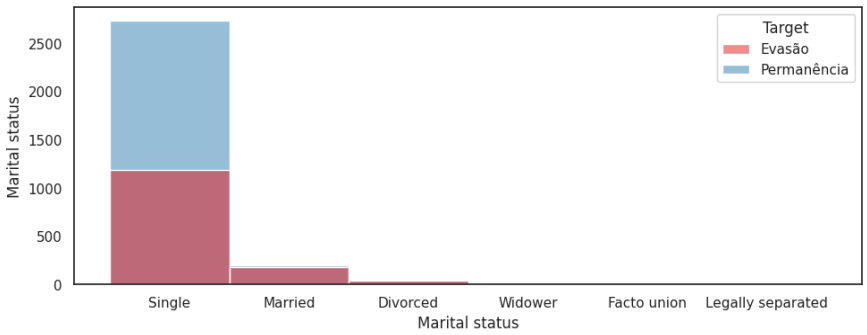

Portuguese
Romanian
Spanish
Brazilian
Santomean
Ukrainian
Dutch
Mozambican
Angolan
Mexican
Italian
Cape Verdean
Turkish
Moldova (Republic of)
Guinean
Colombian
German
Cuban
Russian
English
Lithuanian
Nacionality
Target
Evasão
Permanência
0
500
1000
1500
2000
2500
3000
Count

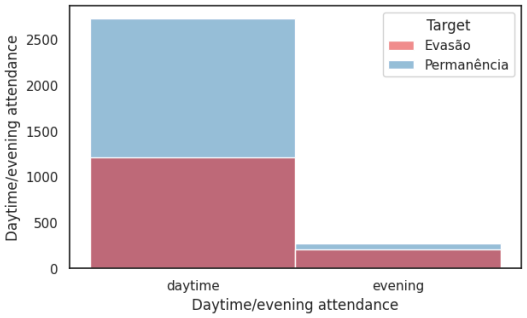

Mother's qualification
General commerce course
Secondary Education—12th Year of Schooling or Equivalent
General Course of Administration and Commerce
Supplementary Accounting and Administration
Higher Education—degree
Higher Education—master's degree
Basic education 1st cycle (4th/5th year) or equivalent
Higher Education—bachelor's degree
2nd cycle of the general high school course
Other—11th Year of Schooling
Cannot read or write
12th Year of Schooling—not completed
Higher Education—doctorate
Unknown
7th Year (Old)
Can read without having a 4th year of schooling
7th year of schooling
2nd year complementary high school course
9th Year of Schooling—not completed
8th year of schooling
Frequency of Higher Education
11th Year of Schooling—not completed
Complementary High School Course—not concluded
Basic Education 2nd Cycle (6th/7th/8th Year) or equivalent
10th Year of Schooling
Basic Education 3rd Cycle (9th/10th/11th Year) or Equivalent
Technical-professional course
Complementary High School Course
Technological specialization course
Target
Evasão
Permanência
0
100
200
300
400
500
600
700
800
Count

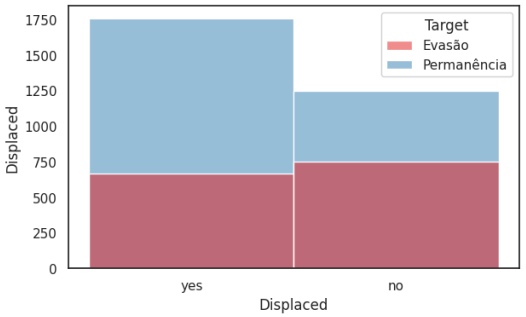

Target
Evasão
Permanência
Other—11th Year of Schooling
Higher Education—degree
Basic education 1st cycle (4th/5th year) or equivalent
Basic Education 2nd Cycle (6th/7th/8th Year) or equivalent
Secondary Education—12th Year of Schooling or Equivalent
Basic Education 3rd Cycle (9th/10th/11th Year) or Equivalent
Higher Education—doctorate
Higher Education—master's degree
Unknown
Higher Education—bachelor's degree
Technological specialization course
7th Year (Old)
12th Year of Schooling—not completed
Can read without having a 4th year of schooling
7th year of schooling
Higher education—degree (1st cycle)
10th Year of Schooling
Complementary High School Course
Cannot read or write
Specialized higher studies course
Technical-professional course
2nd year complementary high school course
9th Year of Schooling—not completed
Higher Education—master's degree (2nd cycle)
General commerce course
Professional higher technical course
11th Year of Schooling—not completed
Frequency of Higher Education
8th year of schooling
Complementary High School Course—not concluded
Higher Education—doctorate (3rd cycle)
Supplementary Accounting and Administration
2nd cycle of the general high school course
General Course of Administration and Commerce
Father's qualification
0
100
200
300
400
500
600
700
800
Count

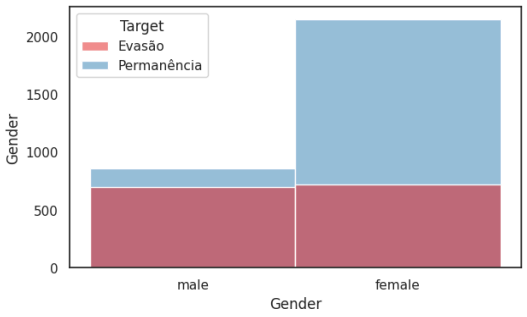

Target
Evasão
Permanência
Personal Services, Security and Safety Workers, and Sellers
Intermediate Level Technicians and Professions
Unskilled Workers
Skilled Workers in Industry, Construction, and Craftsmen
Administrative staff
Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers
Other Armed Forces personnel
Student
Farmers and Skilled Workers in Agriculture, Fisheries, and Forestry
Specialists in Intellectual and Scientific Activities
Other Situation
Installation and Machine Operators and Assembly Workers
Health professionals
Data, accounting, statistical, financial services, and registry-related operators
(blank)
Other administrative support staff
Intermediate level science and engineering technicians and professions
Personal care workers and the like
Personal service workers
Hotel, catering, trade, and other services directors
Technicians and professionals of intermediate level of health
Specialists in the physical sciences, mathematics, engineering, and related techniques
Armed Forces Professions
Teachers
Armed Forces Sergeants
Office workers, secretaries in general, and data processing operators
Sellers
Armed Forces Officers
Specialists in finance, accounting, administrative organization, and public and commercial relations
Directors of administrative and commercial services
Information and communication technology technicians
Intermediate level technicians from legal, social, sports, cultural, and similar services
Mother's occupation
0
200
400
600
800
1000
Count

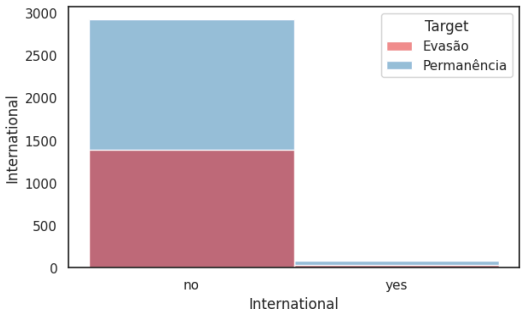

Target
Evasão
Permanência
Unskilled Workers
Intermediate Level Technicians and Professions
Skilled Workers in Industry, Construction, and Craftsmen
Armed Forces Professions
Personal Services, Security and Safety Workers, and Sellers
Installation and Machine Operators and Assembly Workers
Administrative staff
Representatives of the Legislative Power and Executive Bodies, Directors, Directors and Executive Managers
Specialists in Intellectual and Scientific Activities
Specialists in finance, accounting, administrative organization, and public and commercial relations
Farmers and Skilled Workers in Agriculture, Fisheries, and Forestry
Student
Other Situation
Workers in food processing, woodworking, and clothing and other industries and crafts
Specialists in the physical sciences, mathematics, engineering, and related techniques
(blank)
Other administrative support staff
Street vendors (except food) and street service providers
Unskilled workers in agriculture, animal production, and fisheries and forestry
Market-oriented farmers and skilled agricultural and animal production workers
Unskilled workers in extractive industry, construction, manufacturing, and transport
Personal service workers
Assembly workers
Technicians and professionals of intermediate level of health
Intermediate level science and engineering technicians and professions
Meal preparation assistants
Farmers, livestock keepers, fishermen, hunters and gatherers, and subsistence
Information and communication technology technicians
Data, accounting, statistical, financial services, and registry-related operators
Skilled construction workers and the like, except electricians
Other Armed Forces personnel
Skilled workers in metallurgy, metalworking, and similar
Sellers
Vehicle drivers and mobile equipment operators
Health professionals
Armed Forces Sergeants
Fixed plant and machine operators
Intermediate level technicians from legal, social, sports, cultural, and similar services
Teachers
Directors of administrative and commercial services
Personal care workers and the like
Skilled workers in electricity and electronics
Office workers, secretaries in general, and data processing operators
Hotel, catering, trade, and other services directors
Armed Forces Officers
Protection and security services personnel
Father's occupation
0
100
200
300
400
500
600
700
Count

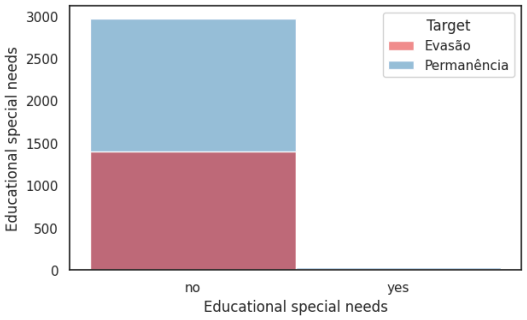

Target
Evasão
Permanência
Marital status
0
500
1000
1500
2000
2500
Single
Married
Divorced
Widower
Facto union
Legally separated
Marital status

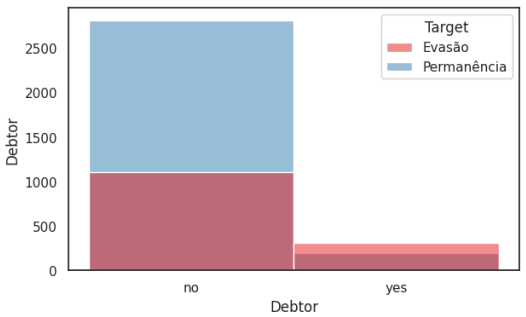

Target
Evasão
Permanência
Daytime/evening attendance
0
500
1000
1500
2000
2500
daytime
evening
Daytime/evening attendance

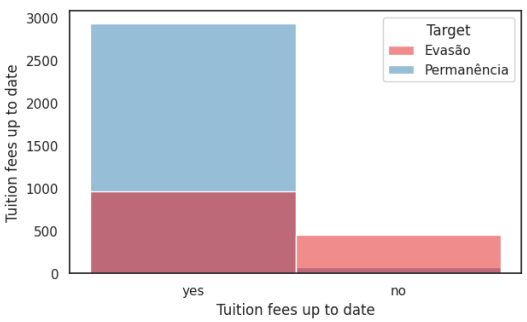

Target
Evasão
Permanência
Displaced
0
250
500
750
1000
1250
1500
1750
yes
no
Displaced

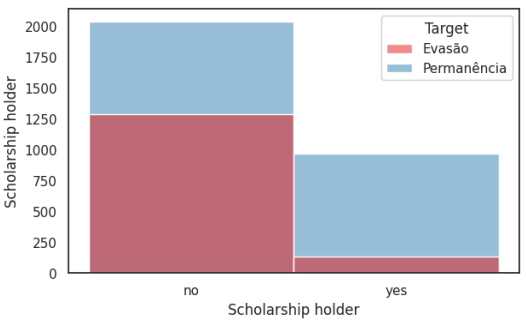

Target
Evasão
Permanência
Gender
0
500
1000
1500
2000
male
female
Gender


Target
Evasão
Permanência
International
0
500
1000
1500
2000
2500
3000
no
yes
International


Target
Evasão
Permanência
Educational special needs
0
500
1000
1500
2000
2500
3000
no
yes
Educational special needs


Target
Evasão
Permanência
Debtor
0
500
1000
1500
2000
2500
no
yes
Debtor


Target
Evasão
Permanência
Tuition fees up to date
0
500
1000
1500
2000
2500
3000
yes
no
Tuition fees up to date


Target
Evasão
Permanência
Scholarship holder
0
250
500
750
1000
1250
1500
1750
2000
no
yes
Scholarship holder