Instituto Politécnico do Cávado e do Ave

Escola Superior de Tecnologia

Licenciatura

em

Engenharia Informática Médica

Inteligência Artificial

Bruno Rafael Mendes Oliveira – a15566

Diogo Mário Sá Fernandes – a24017

Janeiro de 2024

Esta página foi deixada em branco propositadamente.

Instituto Politécnico do Cávado e do Ave

Escola Superior de Tecnologia

Licenciatura

em

Engenharia Informática Médica

Relatório do Projeto Engenharia de Software

Previsão de Abandono no Ensino Superior Usando *Machine Learning*

Unidade Curricular

Inteligência Artificial

Nome dos Alunos

Bruno Oliveira

Diogo Fernandes

Docente da Unidade Curricular:

Profª. Joaquim Gonçalves

Dezembro de 2023

Esta página foi deixada em branco propositadamente.

Resumo

Este relatório documenta o estudo realizado na unidade curricular de Inteligência Artificial, focada na pesquisa de previsão de desistência no ensino superior usando técnicas avançadas de M*achine Learning*. O objetivo principal foi aplicar os conceitos teóricos da disciplina para criar um modelo capaz de identificar alunos em risco de abandonar suas instituições educacionais. Durante esse projeto, foi realizada uma análise minuciosa dos dados dos alunos, incluindo a preparação prévia, uma pesquisa detalhada e o equilíbrio das classes.

As técnicas que foram utilizadas envolveram a normalização dos dados, remover variáveis com baixa correlação em relação ao objetivo e usar métodos de reamostragem como o *SMOTE* para lidar com o desequilíbrio das classes. Eu avaliei e otimizei diversos modelos de aprendizado de máquina, como Regressão Logística, *Random Forest, XGBoost* e Redes Neurais, ajustando seus Hiper Parâmetros.

Os resultados obtidos destacam a eficácia das técnicas de inteligência artificial na previsão da desistência escolar e ressaltam a importância das abordagens sofisticadas e complexas de ensinar uma máquina no campo educacional a calcular/classificar alguma decisão. Esse trabalho não apenas me permitiu aplicar os conhecimentos teóricos adquiridos na prática, mas também enfatizou o papel crucial da análise dos dados e da modelagem preditiva na solução de problemas reais.

Palavras-Chave: Inteligência Artificial, *Machine Learning*, Previsão de Abandono Escolar, Análise de Dados, Modelagem Preditiva, *Python*.

Abstract

This report documents the study conducted in the Artificial Intelligence course, focused on research into predicting dropout in higher education using advanced Machine Learning techniques. The main objective was to apply the theoretical concepts of the course to create a model capable of identifying students at risk of dropping out of their educational institutions. During this project, a thorough analysis of student data was carried out, including prior preparation, detailed research, and class balancing.

The techniques used involved normalizing the data, removing variables with low correlation to the target, and using resampling methods like SMOTE to deal with class imbalances. I evaluated and optimized various machine learning models such as Logistic Regression, Random Forest, XGBoost, and Neural Networks by adjusting their Hyperparameters.

The results obtained highlight the effectiveness of artificial intelligence techniques in predicting school dropout and emphasize the importance of sophisticated and complex approaches to teaching a machine in the educational field to calculate/predict a decision. This work not only allowed me to apply the theoretical knowledge acquired in practice but also emphasized the crucial role of data analysis and predictive modeling in solving real problems.

Keywords: *Artificial Intelligence, Machine Learning, Prediction of School Dropout, Data Analysis, Predictive Modeling, Python.*

Índice

Índice de Figuras

Índice de Tabelas

Lista de siglas e acrónimos

- *AI*: *Artificial Inteligence*;
- UC: Unidade Curricular
- XGBoost: eXtreme Gradient Boosting

1. Introdução

1.1. Enquadramento

No contexto da Unidade Curricular (UC) de Inteligência Artificial, aprofundar o conhecimento em *Machine Learning* é essencial para o desenvolvimento de sistemas inteligentes capazes de tomar decisões complexas e proporcionar soluções inovadoras para problemas do mundo real. O *Machine learning*, um ramo vital da Inteligência Artificial, envolve o desenvolvimento de algoritmos que permitem que as máquinas aprendam e façam previsões ou decisões com base em dados.

Este projeto concentra-se na aplicação de técnicas avançadas de *Machine Learning* para prever a desistência de alunos no ensino superior, um desafio significativo que enfrentam muitas instituições educacionais. A capacidade de prever com precisão quais alunos estão em risco de abandonar seus estudos permite intervenções oportunas, melhorando assim as taxas de retenção e o sucesso educacional.

1.2. Objetivos

Os principais objetivos deste projeto são:

- **Desenvolver e Avaliar Modelos de *Machine Learning*:** Implementar e avaliar diversos modelos de aprendizado de máquina, como Regressão Logística, Random Forest, XGBoost e Redes Neurais, para prever o abandono escolar.
- **Otimização de Modelos através de Técnicas Avançadas:** Utilizar técnicas como normalização de dados, balanceamento de classes e ou ajuste de Hiper parâmetros para melhorar o desempenho dos modelos.
- **Análise Exploratória e Preparação de Dados:** Realizar uma análise exploratória para compreender as características dos dados e aplicar métodos de pré-processamento para preparar os dados para o treinamento de modelos.
- **Comparação de Desempenho e Seleção de Modelo:** Comparar o desempenho dos diferentes modelos com base em métricas como Accuracy, Precision, Recall e F1-Score para selecionar o modelo mais eficaz.
- **Aplicação Prática em Contexto Educacional:** Explorar a aplicabilidade prática do modelo no contexto educacional, fornecendo insights sobre como as instituições de ensino superior podem utilizar essas previsões para reduzir as taxas de abandono.

2. Metodologia

Os dados fornecidos para o projeto são cruciais para entender o fenômeno do abandono no ensino superior. Os registos contêm variadas informações sobre os alunos, potencialmente oferecendo *insights* sobre fatores que influenciam a decisão de continuar ou desistir dos estudos.

O conjunto de dados é composto por 5411 registos, cada um representando um aluno, onde 651 são casos de alunos que abandonaram o ensino superior e 4761 são os que não abandonaram o ensino superior. Existem 34 atributos distintos onde estes atributos abrangem uma ampla gama de fatores, incluindo dados institucionais, cursos, informações demográficas, como gênero e idade, bem como informações socioeconômicas e acadêmicas.

Uma inspeção inicial indica que o conjunto de dados está completo, sem valores ausentes em nenhum dos atributos. Este é um aspeto positivo que facilita a análise e o treino dos modelos, eliminando a necessidade de etapas adicionais de imputação de dados.

A análise descritiva revelou uma variedade de padrões. A correlação entre os atributos e a variável de abandono varia, indicando que certas características podem ter mais influência sobre a decisão do aluno de abandonar os estudos.

A matriz de correlação fornece uma base para a seleção de variáveis, com a exclusão daquelas que possuem baixa correlação com a variável alvo, evitando redundâncias e reduzindo a dimensionalidade do modelo.

3. Descrição dos Algoritmos de Machine Learning Utilizados

3.1. Regressão Logística

A Regressão Logística é um algoritmo de classificação estatística que é usado para prever a probabilidade de uma variável dependente categórica. No contexto do projeto, é utilizada para prever a probabilidade de um aluno abandonar a instituição de ensino superior. A simplicidade da Regressão Logística e a sua interoperabilidade fazem dela uma excelente escolha para o modelo base, proporcionando um ponto de referência para a performance dos modelos mais complexos.

3.2. Random Forest

Random Forest é um algoritmo de aprendizado ensemble que constrói múltiplas árvores de decisão durante o treinamento e gera a classe como a moda das classes (classificação) ou previsão média (regressão) das árvores individuais. *Random Forest* é conhecido por sua robustez e capacidade de lidar com conjuntos de dados com um grande número de variáveis, tornando-o apropriado para analisar os fatores que podem influenciar o abandono escolar.

3.3. XGBoost

O XGBoost (*eXtreme Gradient Boosting*) é uma implementação otimizada de árvores de decisão com *boosting* de gradiente projetada para velocidade e performance. É um dos algoritmos mais eficazes devido à sua velocidade e desempenho. No projeto, o *XGBoost* é utilizado para identificar os estudantes em risco de abandono, tirando vantagem de seu poder preditivo e capacidade de lidar automaticamente com valores ausentes.

4. Desenvolvimento do projeto

O projeto iniciou com uma análise criteriosa de cada variável presente no conjunto de dados fornecido. O objetivo era identificar qualquer atributo que tivesse uma forte correlação com o fenômeno do abandono escolar. À primeira vista todos os atributos pareciam relevantes, era imperativo confirmar sua utilidade e impacto na previsão de abandono.

Com um entendimento firme sobre o conjunto de dados, a pesquisa focou na seleção de algoritmos de *machine learning* adequados que pudessem fornecer *insights* confiáveis. Após uma avaliação, os seguintes algoritmos foram escolhidos pela sua robustez e adequação ao problema:

- Regressão Logística
- XGBoost
- RandomForest
- Redes Neurais

A Regressão Logística foi o ponto de partida, fornecendo uma compreensão fundamental das necessidades de dados para treino, o processo de treino em si, e a classificação de novas instâncias. Esse conhecimento estabeleceu a base para a aplicação dos demais algoritmos.

O processamento dos dados começou com a leitura do arquivo *.xlsx*, seguido pela exclusão da primeira linha que continha cabeçalhos, irrelevantes para o treino do modelo. Definimos então as variáveis de entrada (características) e a variável alvo, sendo esta última o foco das previsões do nosso modelo.

A divisão das amostras em conjuntos de treino e teste seguiu uma proporção de 90/10 ou 95/5, com o uso consistente de um *random_state* de 42 para assegurar a reprodutibilidade das divisões. Antes de proceder ao treinamento, normalizamos os dados para otimizar o desempenho do algoritmo.

A eficácia dos modelos foi medida através da construção de matrizes de confusão e do cálculo de métricas cruciais como *Accuracy*, Precision, *Recall* e *F1-Score*. Essas métricas nos permitiram não apenas avaliar o desempenho dos modelos de forma quantitativa, mas também comparar a eficácia entre os diferentes algoritmos aplicados.

Após a fase de preparação dos dados, iniciamos o processo de treino do modelo utilizando os algoritmos escolhidos. A primeira execução foi realizada sem alterações nos Hiper parâmetros padrão.

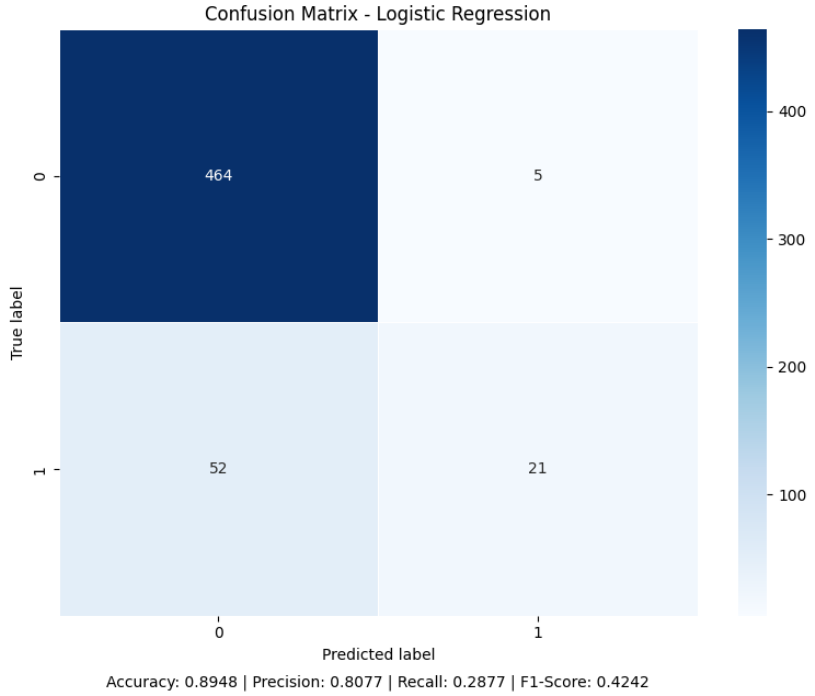

Figura 1 - Matriz de Confusão -*Logist Regression* Primeiro treino – Sem Ajustes de Hyper Parametros

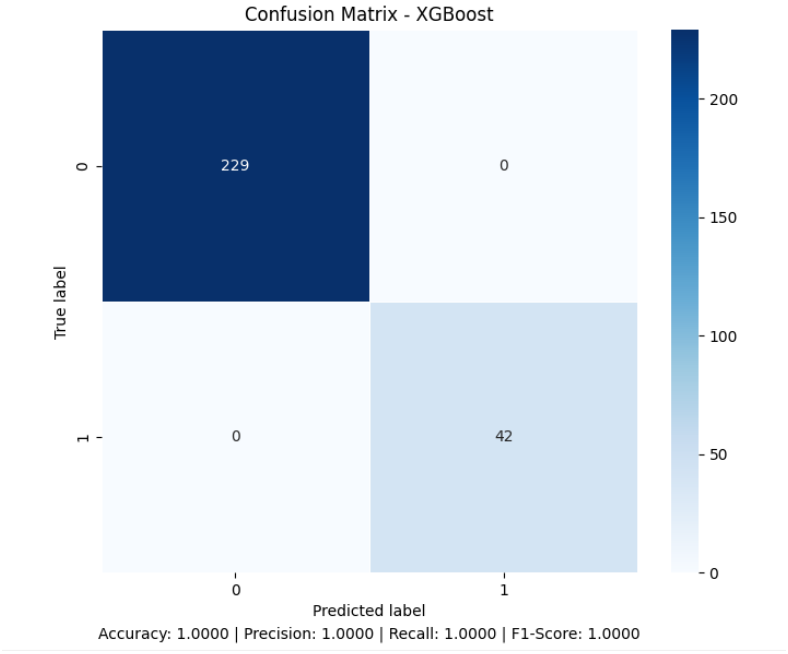

Figura 2 - Matriz de Confusão – *XGBoost* - Primeiro treino – Sem Ajustes de Hyper Parametros

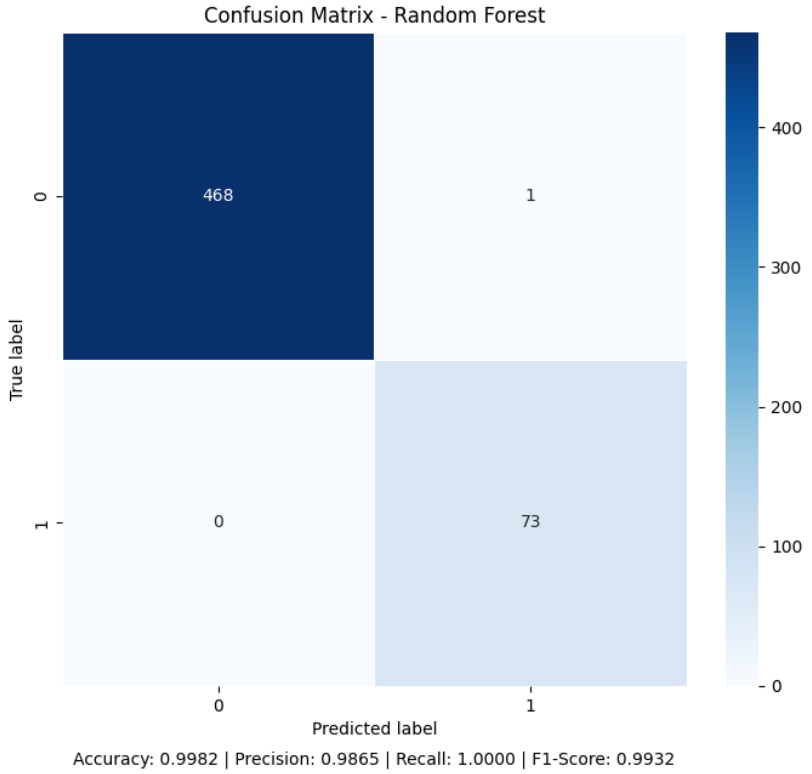

Figura 3 - Matriz de Confusão – *Random Forest* - Primeiro treino – Sem Ajustes de Hyper Parametros

Os resultados iniciais foram considerados satisfatórios, especialmente levando em conta que não houve um esforço de otimização de Hiper parâmetros ou seleção de características. No entanto, apesar desses resultados promissores, estava claro que havia problema ao realizar o modelo.

Então decidimos validar a correlação entre os atributos entre a variável alvo e tambem validar o balanceamento das classes.

Tabela 1 - Balanço entre Classes

Caraterização	Distribuição	Percentagem
0 - Não abandonar	0.881722	88.1722%
1 - Abandonar	0.118278	11.8278%

Aqui validamos que existe uma grande desproporção do *dataset* por isso poderia ser um problema e o modelo estar a entrar em *overfitting*.

Tabela 2 - Correlação entre atributos

Atributo	Correlação
Abandono	1.000000
Id	0.557347
Internacional	0.321051
Ensino Outros Mae	0.318763
Ensino Outros Pai	0.307403
idadeReal	0.076980
sexo	0.076089
nota10-12	0.059191
estado_civil	0.045689
nota12-14	0.022953
cd_regime	0.019851
Ensino Superior Pai	0.010915
1geracao	0.005876
outros	0.005152
trabalhadorEstudante	0.002670
cd_hab_ant	-0.006413
cd_cur_hab_ant	-0.006567
Ensino Superior Mae	-0.009616
maiores 23	-0.011502
cd_instituic	-0.011599
nota18-20	-0.015782
ord_ingresso	-0.020067
Ensino Secundário Pai	-0.024702
Ensino Secundário Mae	-0.026792
nota14-16	-0.030923
pais trabalham	-0.033398
cd_curso	-0.037939
nota16-18	-0.051713
CNA	-0.099619
cd_inst_hab_ant	-0.136754
cd_tip_est_sec	-0.156512

Portugues	-0.291670
Ensino Basico Pai	-0.307403
Ensino Basico Mae	-0.318763

Após uma análise cuidadosa da matriz de correlação, identificamos que a coluna 'Id', que serve unicamente como identificador único para cada aluno, apresentava uma influência desproporcional na variável de interesse. Portanto, optamos por excluir este atributo do conjunto de dados utilizado para treinar o modelo, assegurando que as métricas de desempenho e as previsões geradas refletissem fatores substantivos, não artefactos dos dados.

Tabela 3 - Remover ID das variáveis

Atributo	Correlação
Abandono	1.000000
internacional	0.321051
Ensino Outros Mae	0.318763
Ensino Outros Pai	0.307403
idadeReal	0.076980
sexo	0.076089
nota10-12	0.059191
estado_civil	0.045689
nota12-14	0.022953
cd_regime	0.019851
Ensino Superior Pai	0.010915
1geracao	0.005876
outros	0.005152
trabalhadorEstudante	0.002670
cd_hab_ant	-0.006413
cd_cur_hab_ant	-0.006567
Ensino Superior Mae	-0.009616
maiores 23	-0.011502
cd_instituic	-0.011599
nota18-20	-0.015782
ord_ingresso	-0.020067
Ensino Secundário Pai	-0.024702
Ensino Secundário Mae	-0.026792

nota14-16	-0.030923
pais trabalham	-0.033398
cd_curso	-0.037939
nota16-18	-0.051713
CNA	-0.099619
cd_inst_hab_ant	-0.136754
cd_tip_est_sec	-0.156512
Portugues	-0.291670
Ensino Basico Pai	-0.307403
Ensino Basico Mae	-0.318763

Após remover o atributo Id, observamos uma alteração nas correlações entre os demais atributos, o que era esperado. A exclusão do Id elimina um fator que não possui relevância para a questão do abandono no ensino superior, garantindo que as correlações observadas agora refletem mais adequadamente as relações genuínas entre as variáveis de interesse. Com essa modificação na estrutura dos dados, procedemos a voltar a treinar os modelos, agora com um conjunto de atributos mais representativo dos fatores que influenciam a permanência ou desistência dos alunos no ensino superior.

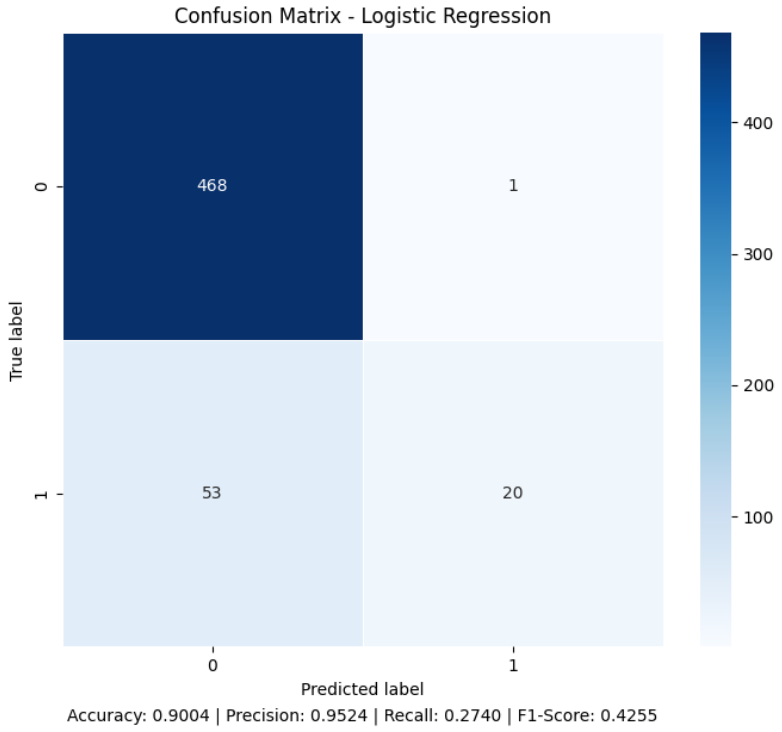

Figura 4 - Matriz de Confusão – *Logistic Regressiont* – Alteração da Corelação

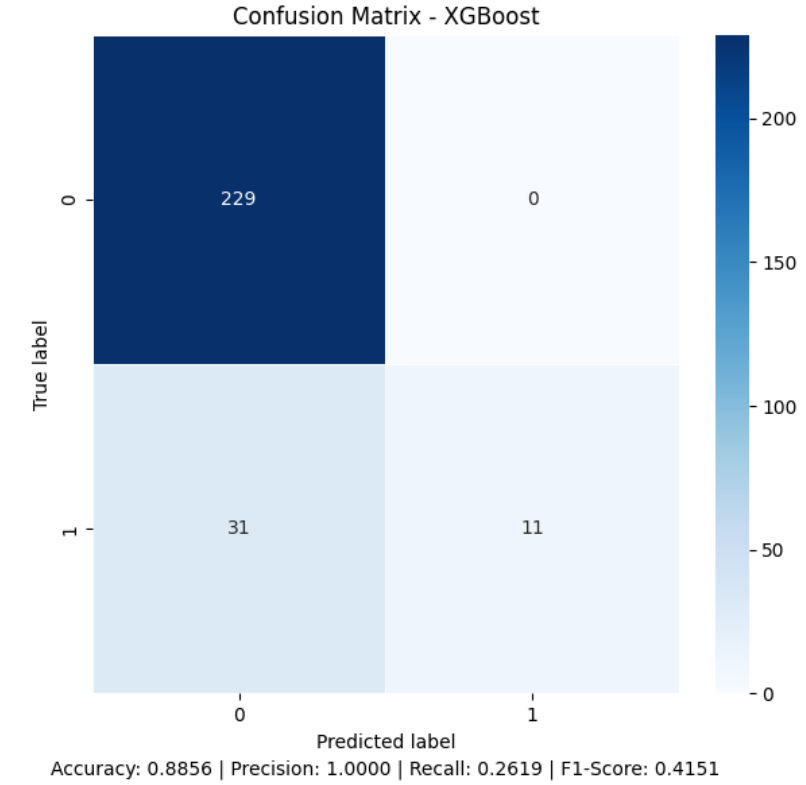

Figura 5 - Matriz de Confusão – *XGBosst* – Alteração da Corelação

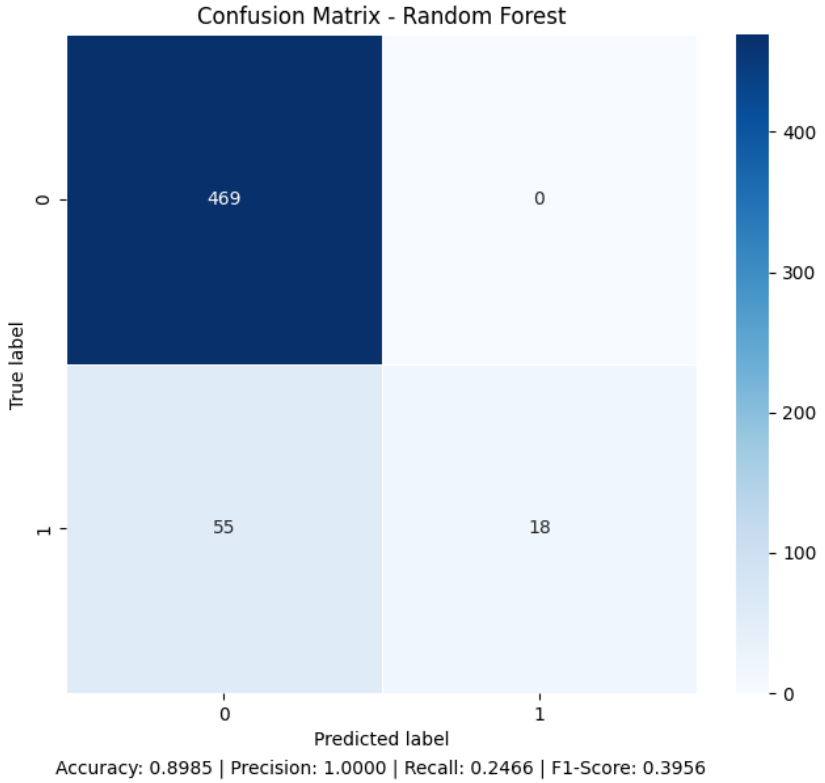

Figura 6 - Matriz de Confusão – *Random Forest* – Alteração da Corelação

Portanto, surgiu a necessidade de equilibrar o conjunto de dados para evitar *overfiting* no treino do modelo decorrente do desequilíbrio entre as classes. Assim, optamos por uma abordagem de amostragem equilibrada, selecionando deliberadamente 500 registos de alunos que não abandonaram o ensino superior e 500 registos de alunos que abandonaram. Esta estratégia visa criar um ambiente de treino mais homogêneo e mais balanceado, no qual cada categoria da variável alvo 'Abandono' é representada igualmente. Com a amostragem balanceada, os resultados obtidos devem proporcionar uma visão mais precisa do impacto dos atributos sobre a probabilidade de abandono do ensino superior.

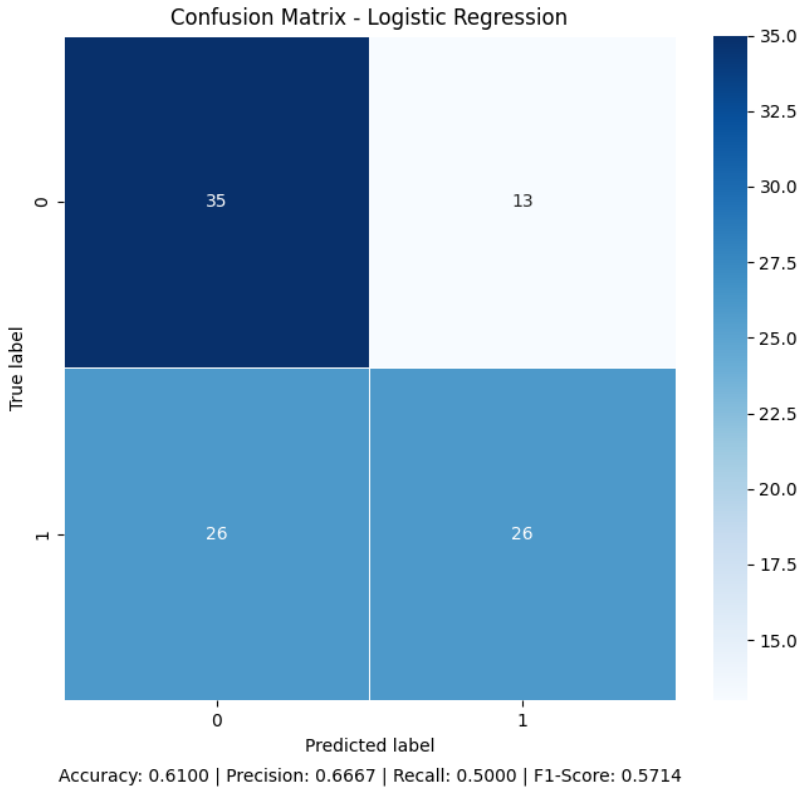

Figura 7 – Matriz de Confusão – *Logistic Regression* – Balanceamento das Classes

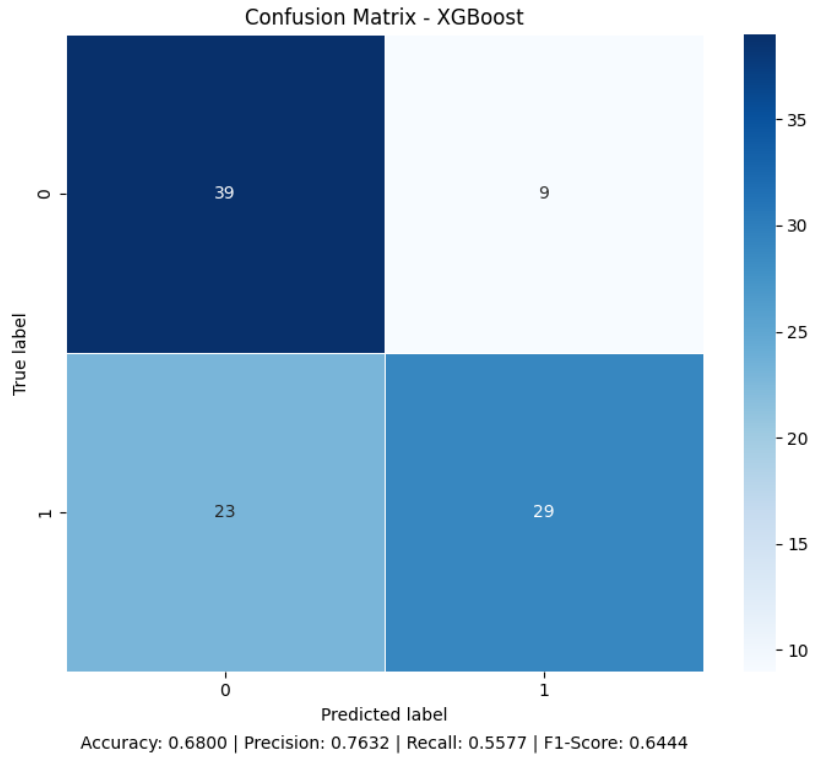

Figura 8 – Matriz de Confusão – *XGBoost* – Balanceamento das Classes

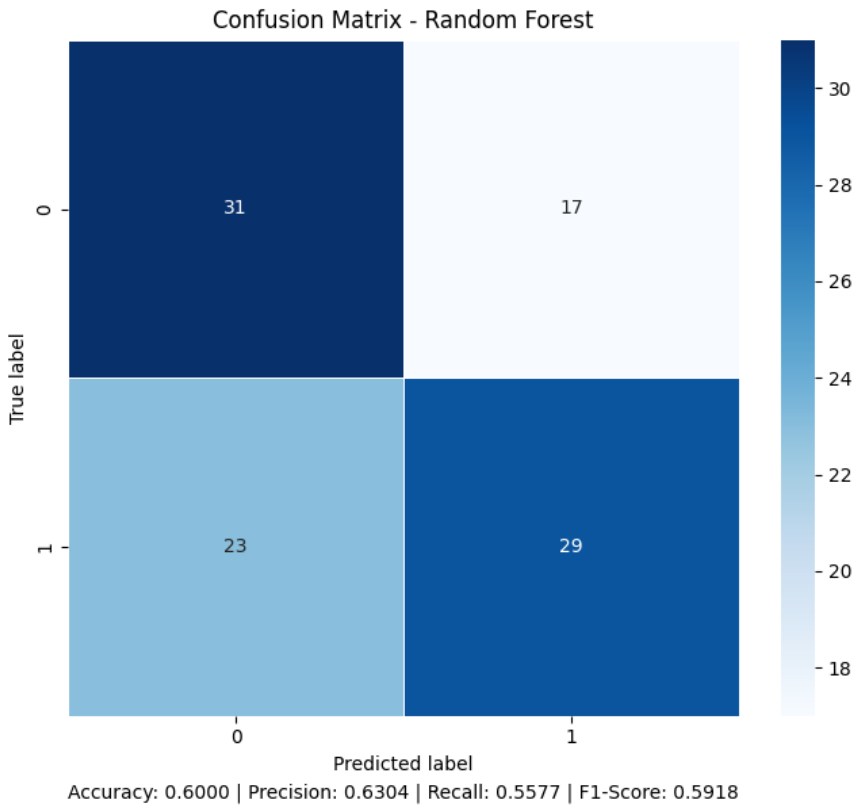

Figura 9 - Matriz de Confusão – *Random Forest* – Balanceamento das Classes

Após balancear o conjunto de dados, observamos uma melhoria percetível na relevância dos resultados. No entanto, uma análise mais aprofundada revelou que, apesar dos indicadores iniciais positivos, nossos modelos estavam incidindo em *overfitting*. Isso significa que, embora os modelos estivessem apresentando uma performance notável na amostra de treino, eles estavam se adaptando excessivamente às especificidades desses dados, comprometendo a capacidade de generalização para novos dados. Esse ajuste excessivo aos dados de treino limita a eficácia dos modelos em prever corretamente o abandono em instâncias não vistas anteriormente, indicando a necessidade de revisão e ajuste nos procedimentos criação do modelo.

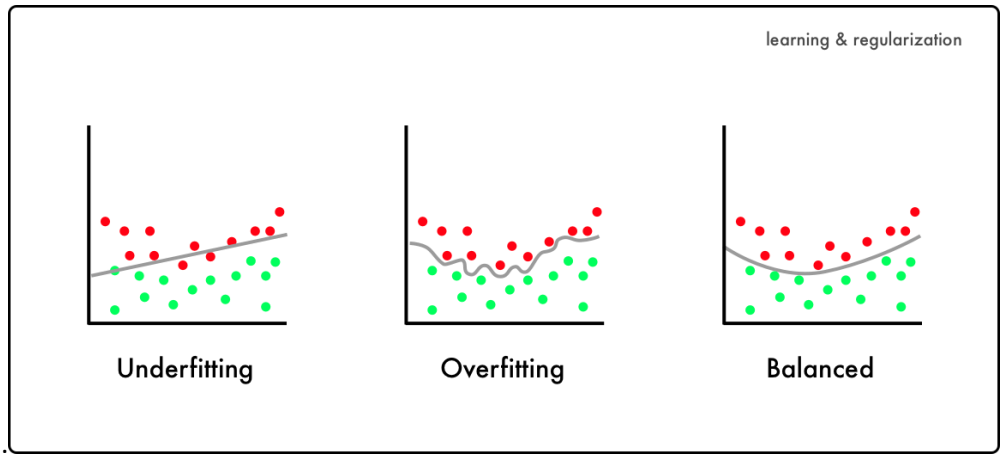

Figura 10 – Dados em *overfiting*

Diante disso, optamos por utilizar gráficos de *learning curves* como ferramenta de diagnóstico para verificar a discrepância observada entre o desempenho nos dados de treino e de validação. Esses gráficos permitiriam uma visualização clara da evolução do desempenho do modelo à medida que aumenta a quantidade de dados de treino, facilitando a identificação de sobre ajuste, ou seja, se o modelo esta a aprender padrões específicos dos dados de treino que não se generalizam bem para dados nunca vistos.

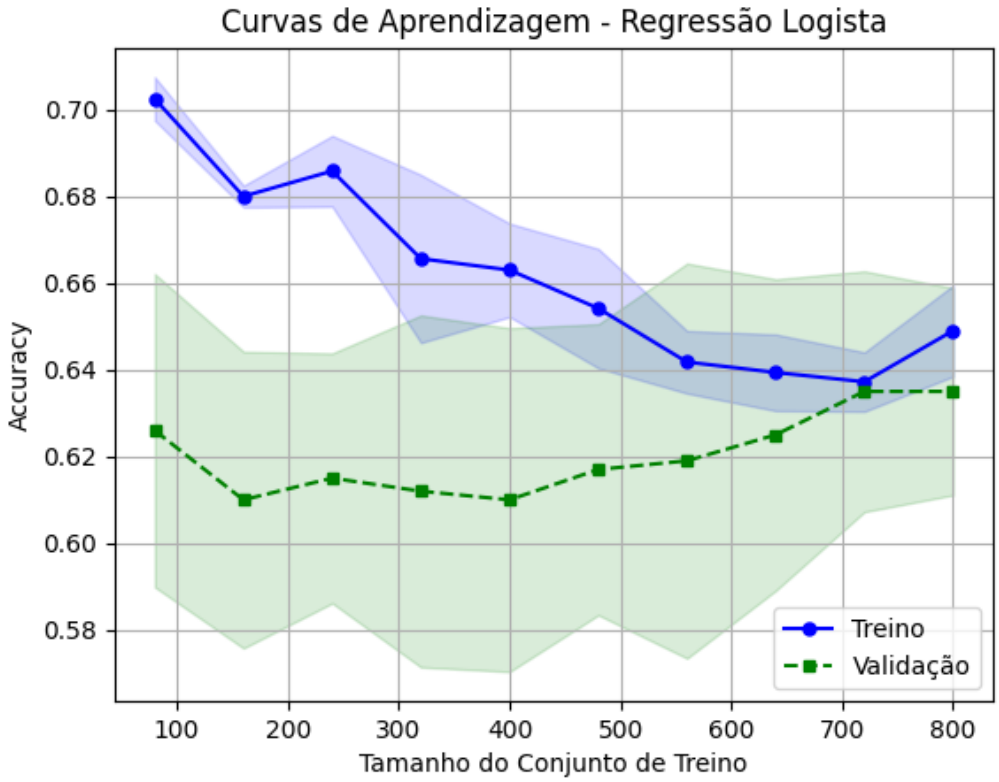

Figura 11 - *Learning Curves - Logistic Regression*

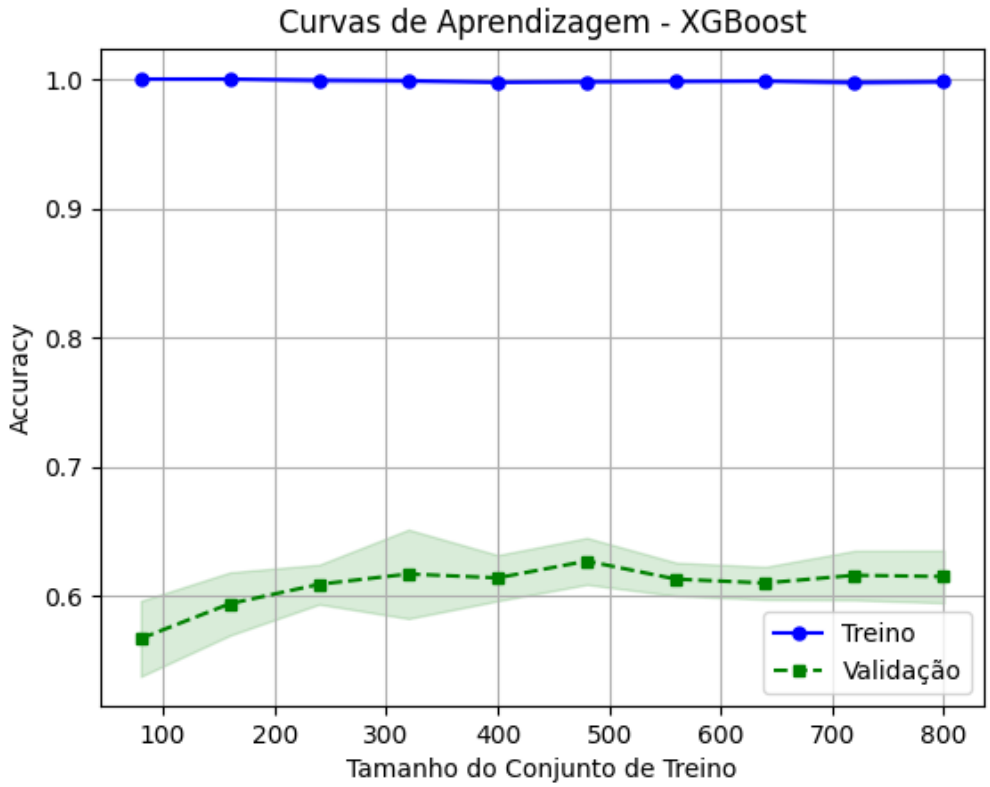

Figura 12 – *Learning Curves - XGBoost*

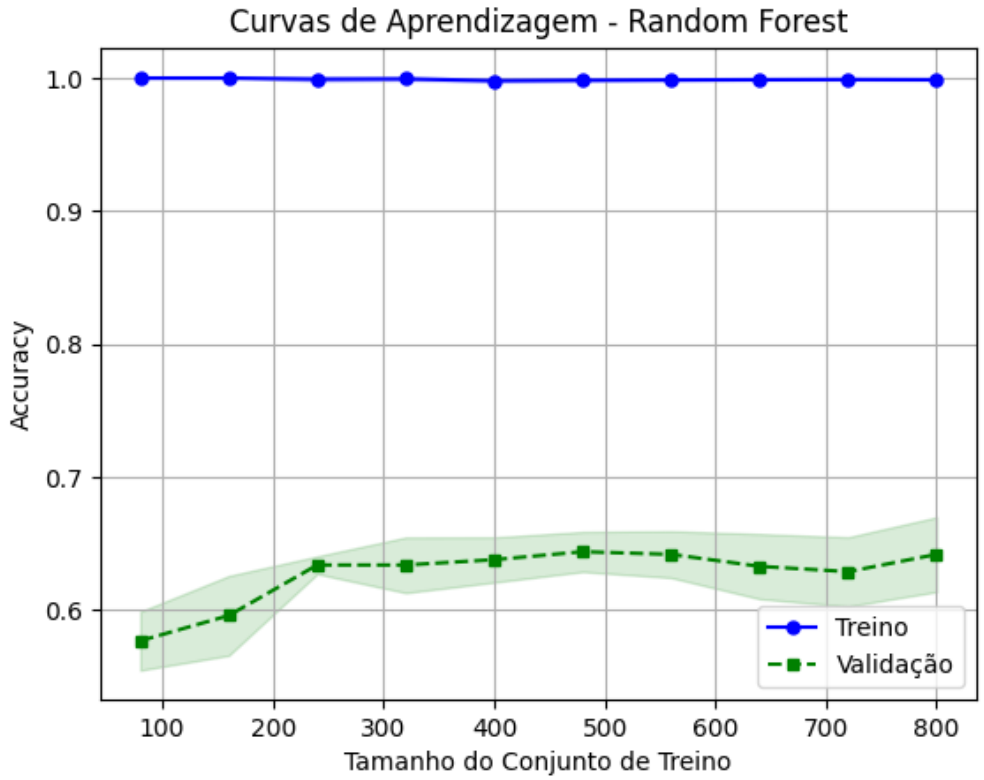

Figura 13 - *Learning Curves - Random Forest*

Com os gráficos de *Learning Curves* foi possível analisar com maior clareza a dinâmica dos modelos. Identificamos uma discrepância significativa entre o desempenho nos conjuntos de treino e de validação, evidenciando que os modelos estavam de fato em *overfitting*. Enquanto os modelos de Random Forest e XGBoost, mostraram uma *accuracy* quase perfeita no treino, não reproduziram o mesmo nível de *accuracy* na validação. Por outro lado, a Regressão Logística, apesar de iniciar com uma *accuracy* mais baixa, apresentou uma melhoria consistente e uma convergência mais próxima entre as curvas de treino e validação, sugerindo uma melhor generalização.

Para mitigar o *overfitting* observado nos modelos e refinar o desempenho geral, iniciamos uma série de experiências focados no ajuste fino dos Hiper parâmetros do treino. Através dessa abordagem iterativa, procurando o equilíbrio ideal entre a capacidade do modelo de obter as nuances dos dados de treino e a sua habilidade de generalizar para dados não vistos, com o objetivo de alcançar um modelo mais robusto e confiável. Os resultados desses ajustes nos forneceram novas perspetivas sobre a relação entre a complexidade do modelo e a sua eficiência de classificação, permitindo-nos avançar em direção a previsões mais precisas e confiáveis sobre o abandono no ensino superior.

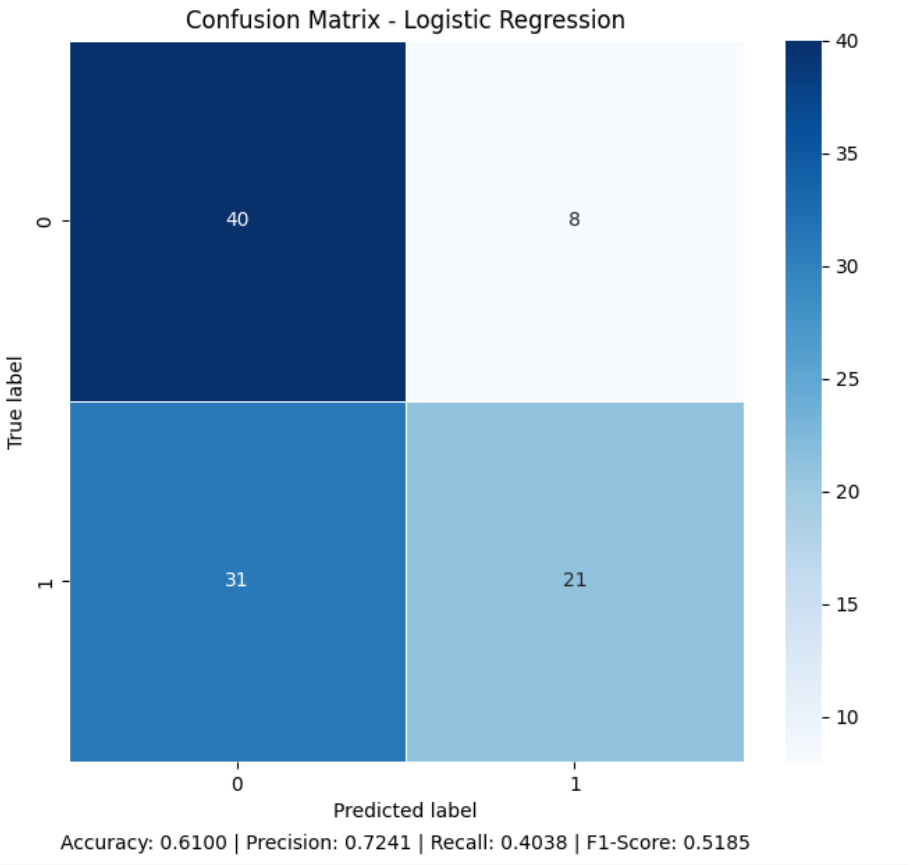

Figura 14 – Matriz de Confusão – Logistic Regression– Ajustes de Hyper Parametros

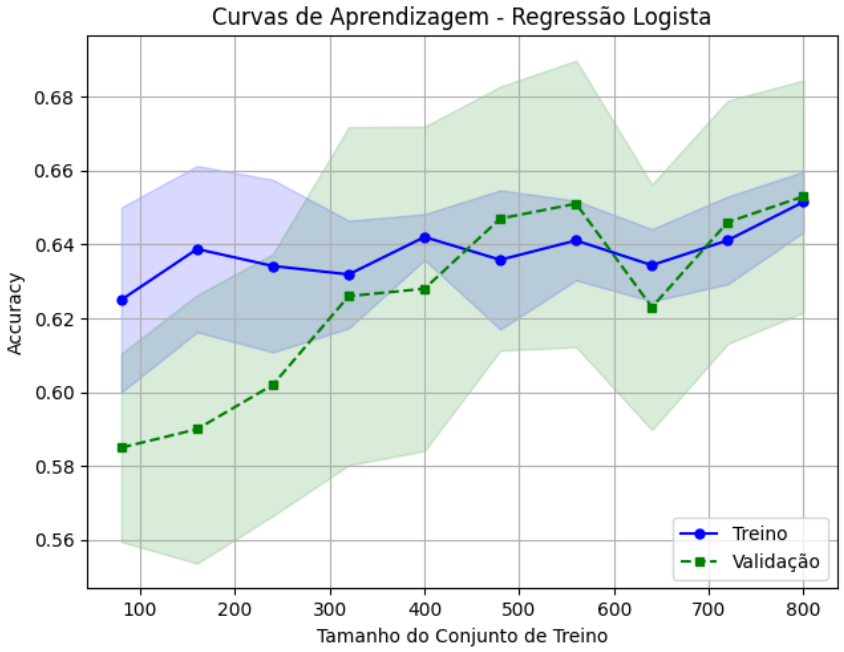

Figura 15 - *Learning Curves - Logistic Regression* – Ajustes de Hyper Parametros

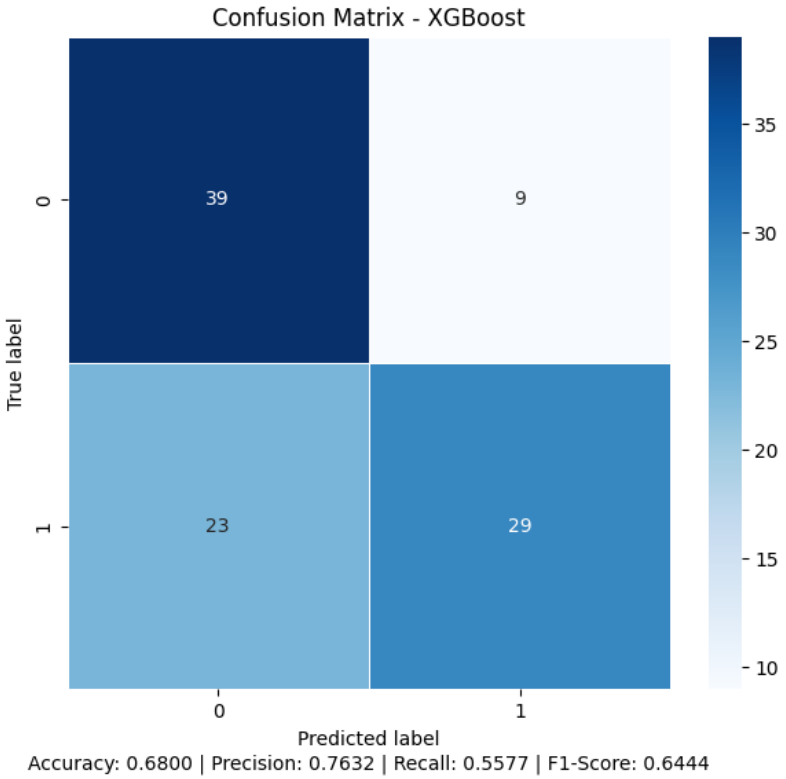

Figura 16 – Matriz de Confusão – *XGBoost*– Ajustes de Hyper Parametros

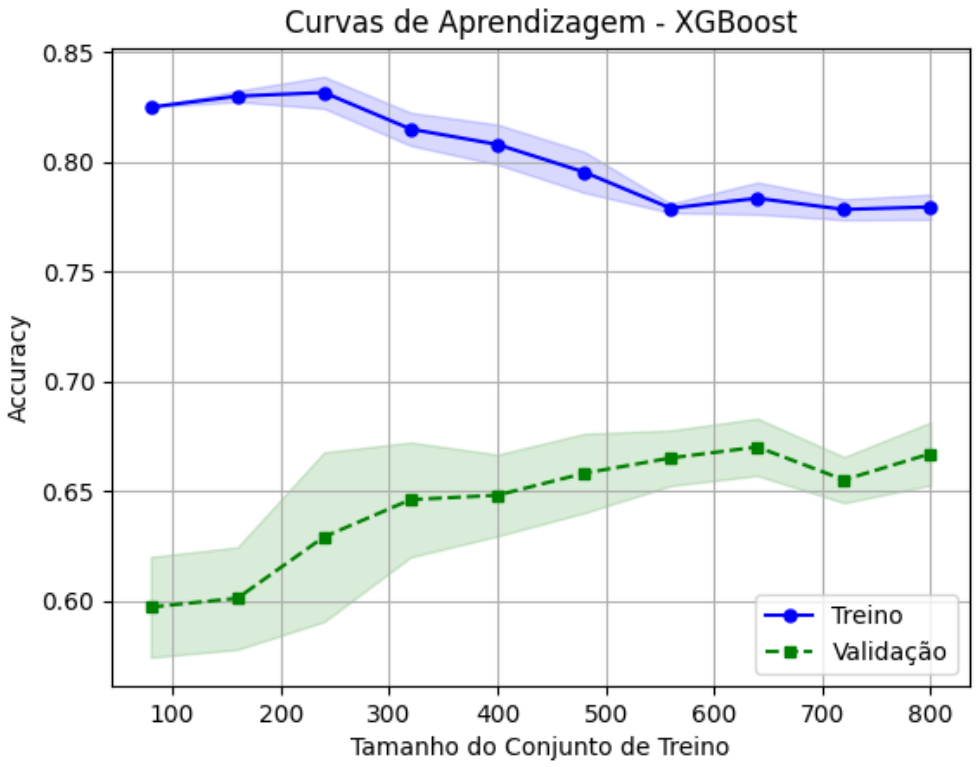

Figura 17 - *Learning Curves - XGBoost* – Ajustes de Hyper Parametros

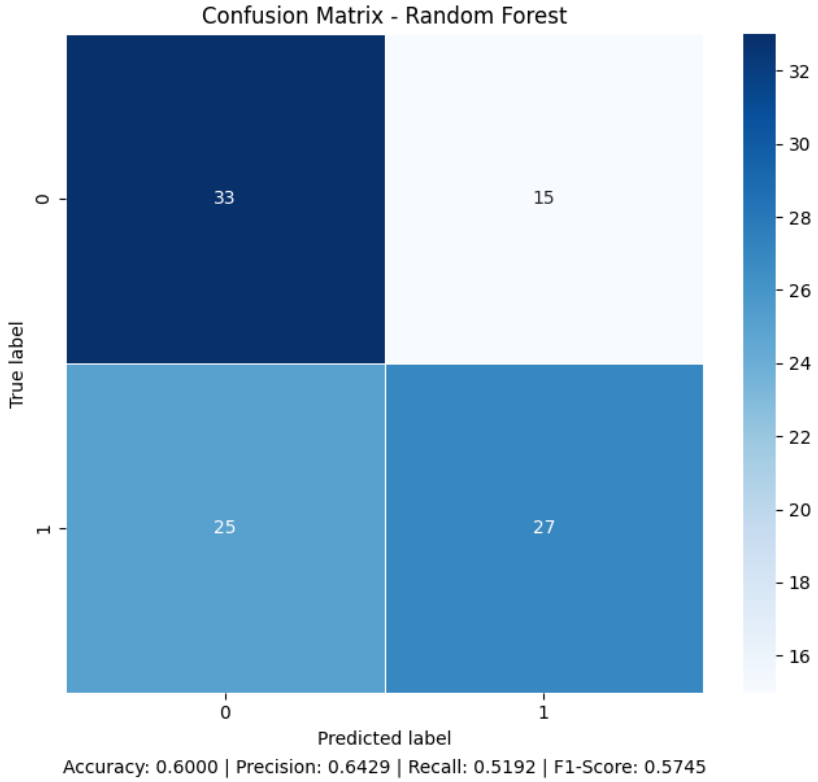

Figura 18 - Matriz de Confusão – *Random Forest* – Ajustes de Hyper Parametros

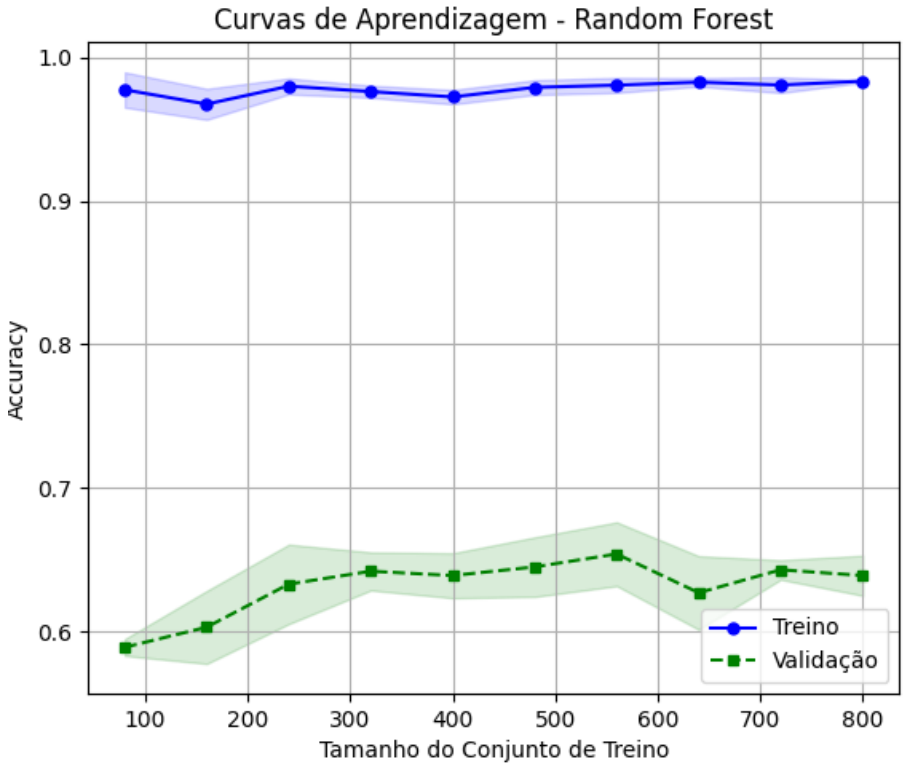

Figura 19 - *Learning Curves – Random Forest* – Ajustes de Hyper Parametros

Após uma análise detalhada das *Learning Curves*, notamos que o modelo de Regressão Logística apresentou avanços significativos a eliminar o *overfitting*, exibindo um aumento gradativo na *accuracy* de validação com o incremento do volume de dados de treino. Tal progresso sugere que as medidas adotadas estão efetivamente aprimorando a capacidade do modelo de generalizar. Por outro lado, os modelos *Random Forest* e *XGBoost*, mantiveram um desempenho relativamente estável, com poucas mudanças entre treino e validação quando comparados às análises anteriores.

Para aprimorar ainda mais os resultados e refinar a precisão dos modelos, voltamos nossa atenção para a influência de cada atributo na variável alvo. Foi realizada uma análise de correlação, que revelou a presença de múltiplos atributos com correlações insignificantes, próximas de zero. Essa constatação levou-nos a concluir que os atributos não tinham contribuição significativa para o poder de classificação do modelo. Portanto, decidimos aplicar um *threshold* de correlação mínima de 0.1 como critério para a exclusão destes atributos do conjunto de dados. Esta estratégia tem como objetivo refinar a estrutura de dados, focando nos atributos com relevância estatística e, consequentemente, potencializando a eficiência do modelo. Neste caso removo-mos as seguintes variaveis ‘idadeReal','sexo','nota10-12','estado_civil','nota12-14', 'cd_regime', '1geracao', 'outros', 'trabalhadorEstudante', 'cd_hab_ant', 'cd_cur_hab_ant', 'maiores 23', 'cd_instituic', 'nota18-20', 'ord_ingresso', 'nota14-16', 'pais trabalham', 'cd_curso', 'nota16-18' ,'CNA'.

Esta etapa de seleção de atributos baseada na correlação é uma prática robusta que não apenas simplifica o modelo, reduzindo sua complexidade, mas também potencializa a interoperabilidade dos resultados. A expectativa é que, ao remover variáveis que contribuem minimamente para a decisão do modelo, possamos obter um ajuste mais preciso e um modelo mais generalizável. Os próximos passos incluem reavaliar as curvas de aprendizagem e o desempenho de validação dos modelos ajustados, para confirmar a eficácia dessa abordagem na melhoria das previsões sobre o abandono no ensino superior.

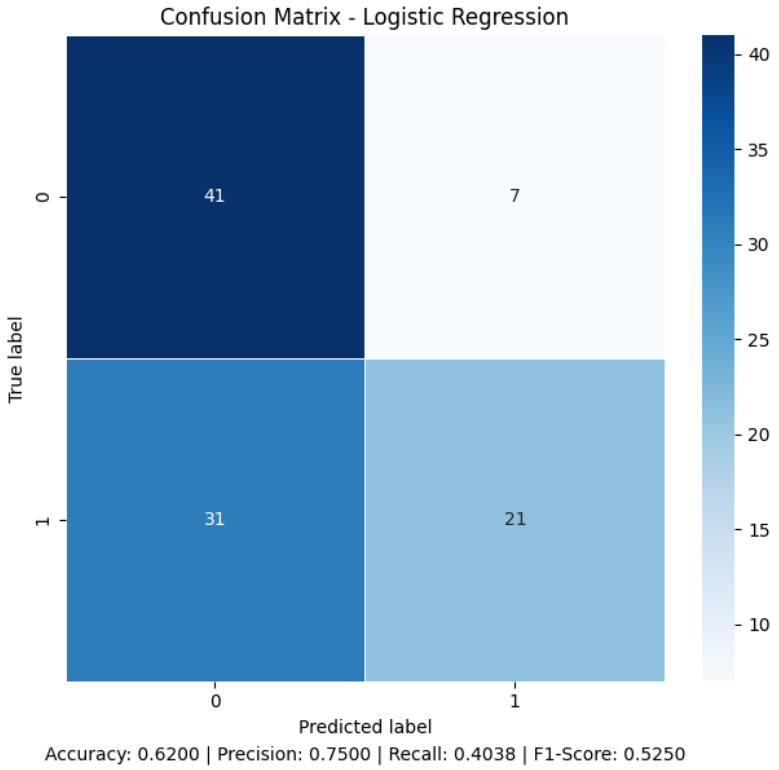

Figura 20 - Matriz de Confusão – *Logistic Regression* – Ajustes de Correlações

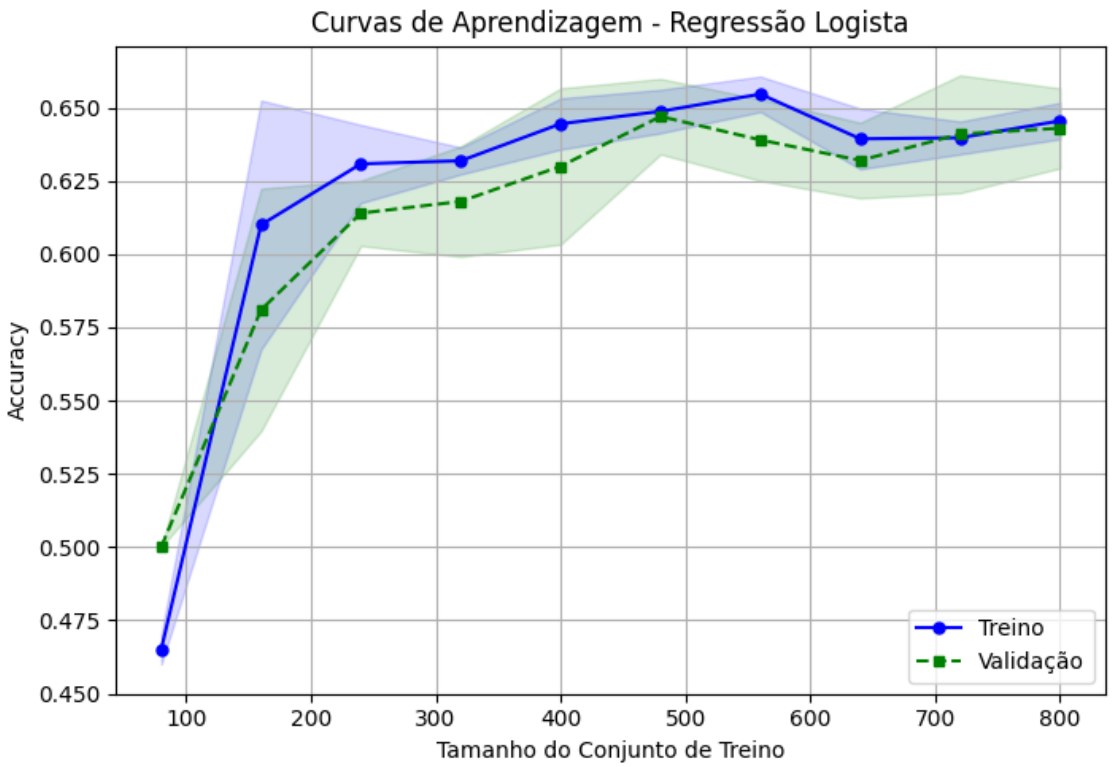

Figura 21 - *Learning Curves – Logistic Regression* – Ajustes de Correlações

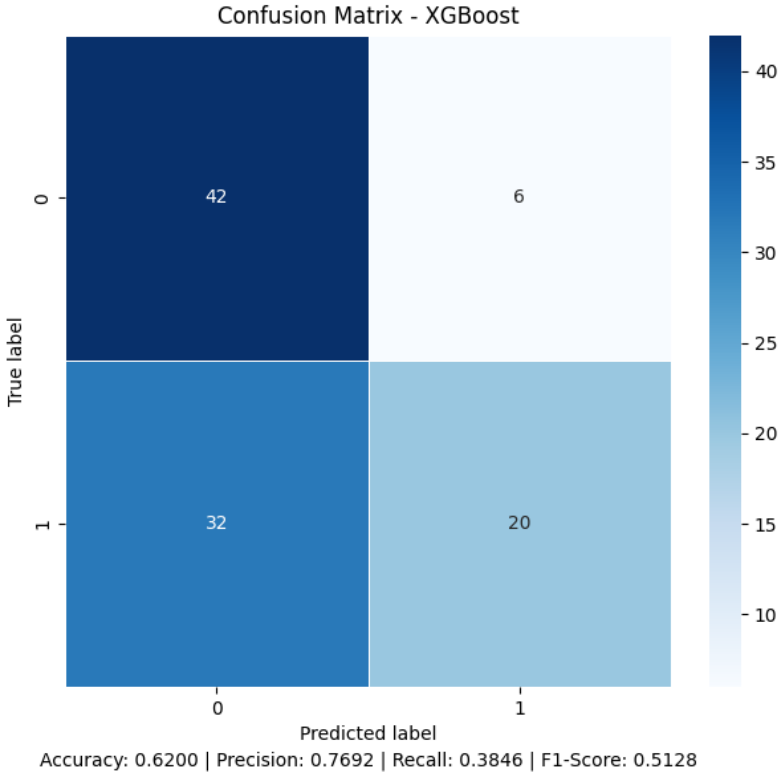

Figura 22 - Matriz de Confusão – *XGBoost* – Ajustes de Correlações

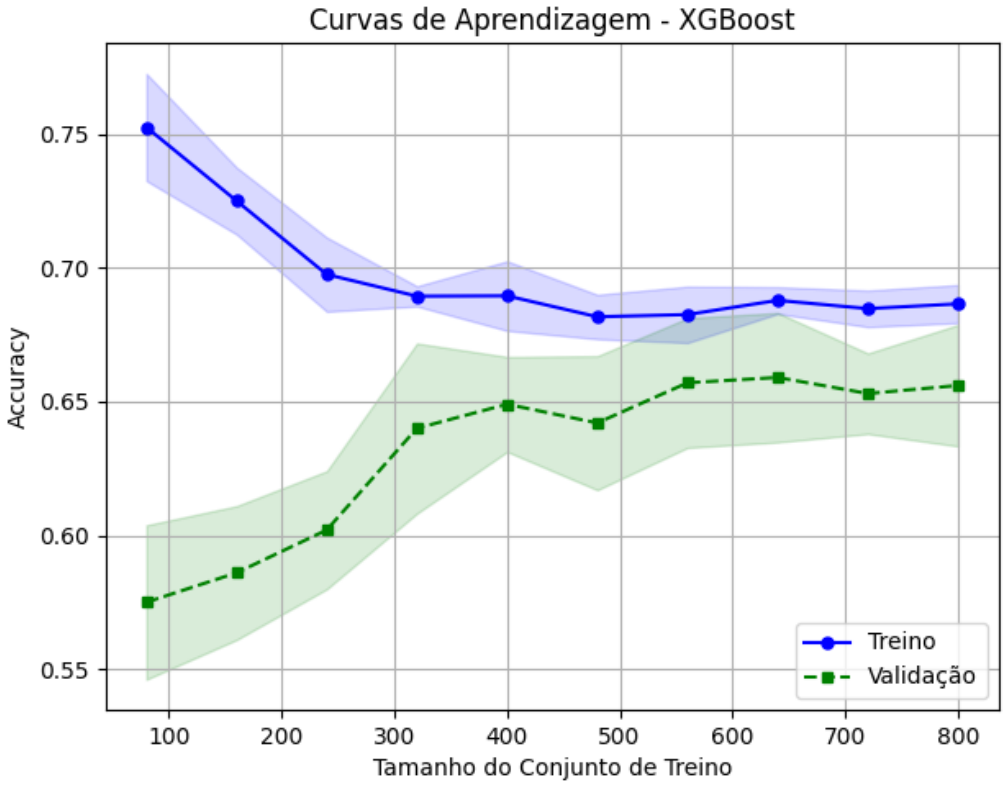

Figura 23 - *Learning Curves – XGBoost* – Ajustes de Correlações

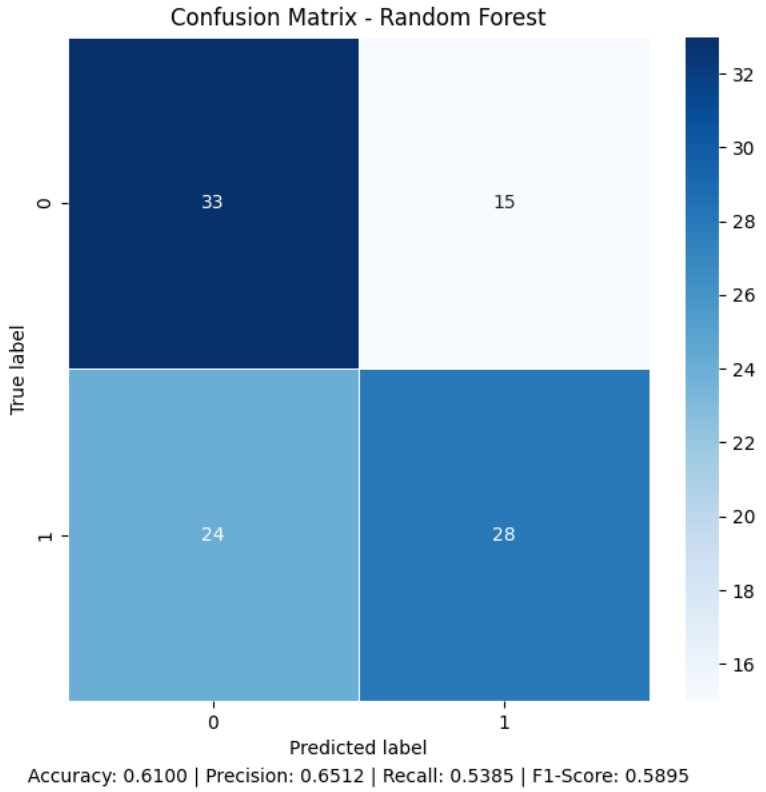

Figura 24 - Matriz de Confusão – *Random Forest* – Ajustes de Correlações

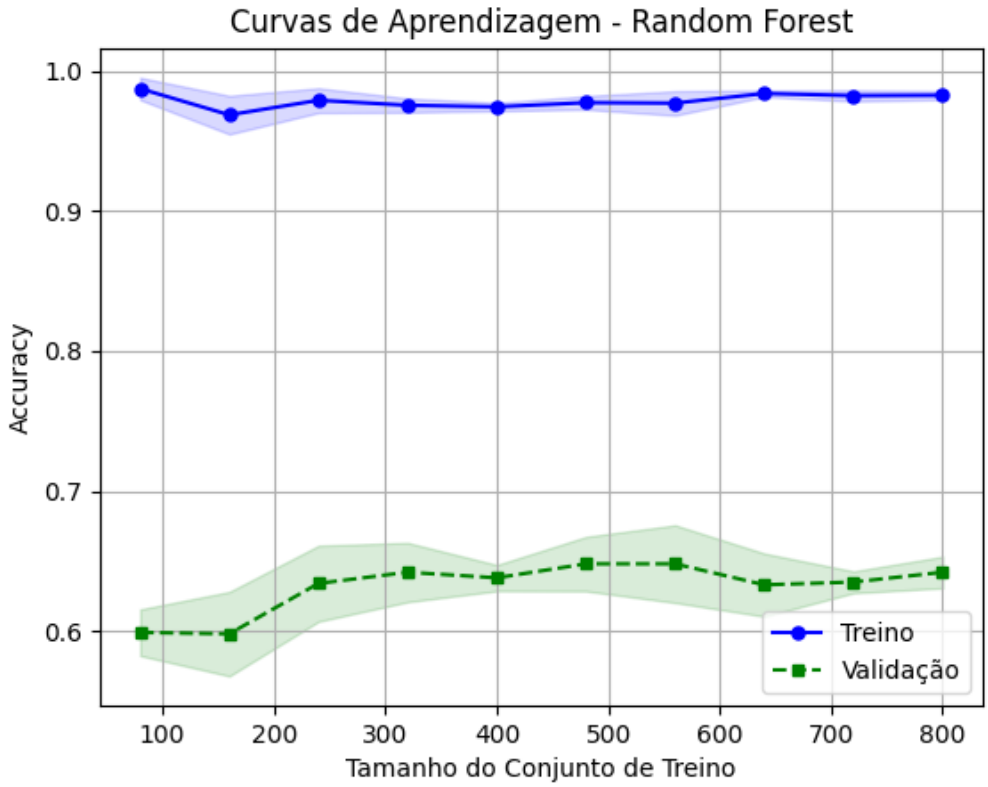

Figura 25 - *Learning Curves – Random Forest* – Ajustes de Correlações

A estratégia de refinar o conjunto de dados ao eliminar atributos com correlações marginais parece ter contribuído para a melhoria dos modelos. Em particular, a Regressão Logística demonstrou uma capacidade aprimorada de generalização, com as curvas de treino e validação se aproximando mais do que antes da remoção dos atributos. Os modelos também mostraram indícios de progresso, embora ainda mostrem sinais de *overfitting*. Esses resultados reforçam a importância de uma seleção cuidadosa de atributos na construção de modelos preditivos e abrem caminho para futuros ajustes e otimizações. A continuidade na revisão dos Hiper parâmetros e a consideração de outras técnicas de regularização e validação cruzada podem oferecer avanços adicionais na busca por modelos robustos e confiáveis.

Este processo de seleção de atributos ajudou a simplificar o modelo, reduzindo o risco de *overfitting* e melhorando a sua generalização. Além disso, a eliminação desses atributos com influência mínima na variável alvo eliminou qualquer potencial de entropia irrelevante que poderia obscurecer as verdadeiras relações sinalizadoras de abandono escolar.

Utilizamos o método de *cross-validation* como medida para verificar a robustez e confiabilidade dos modelos desenvolvidos. Este procedimento desempenha um papel fundamental na avaliação da capacidade dos modelos em generalizar para diferentes conjuntos de dados, reduzindo a influência das particularidades do conjunto inicial. A técnica da *cross-validation* envolve a divisão dos dados em várias partições diferentes, trocando entre conjuntos de treino e teste ao longo das múltiplas iterações. Isso faz com que seja possível utilizar cada observação tanto durante o processo de treino como durante o processo de teste, oferecendo assim uma análise ampla sobre a eficácia do modelo.

Ao usar *cross-validation*, não só aumentamos a confiança nas métricas de desempenho, como também ganhamos *insights* valiosos sobre a estabilidade e a consistência dos modelos em diferentes amostras dos dados. Este passo é crucial para o refinamento final dos modelos, guiando ajustes adicionais nos Hiper parâmetros e na seleção de características, culminando em um sistema de previsão mais preciso e generalizável.

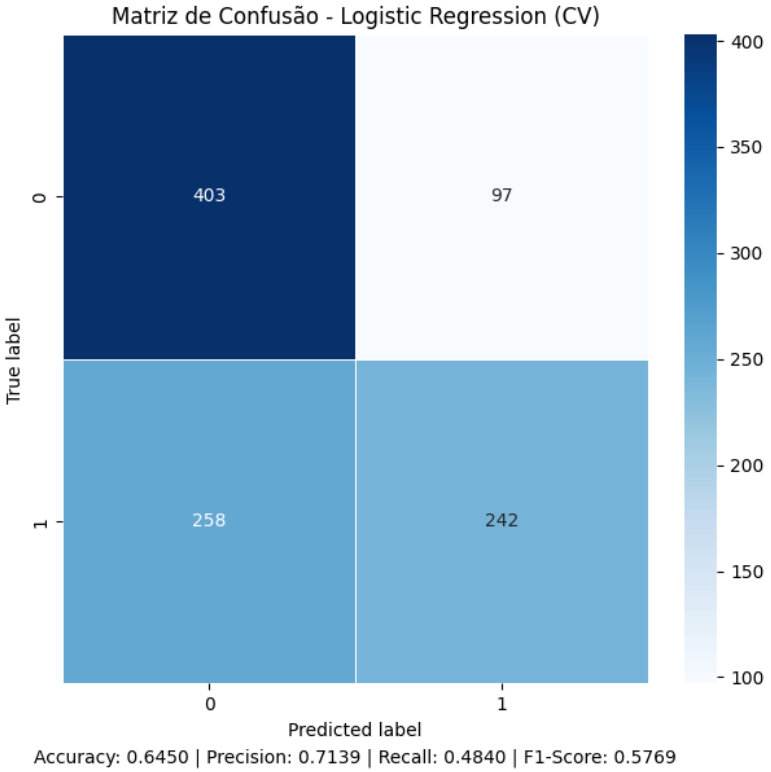

Figura 26 - Matriz de Confusão – *Logistic Regression – Cross-Validation*

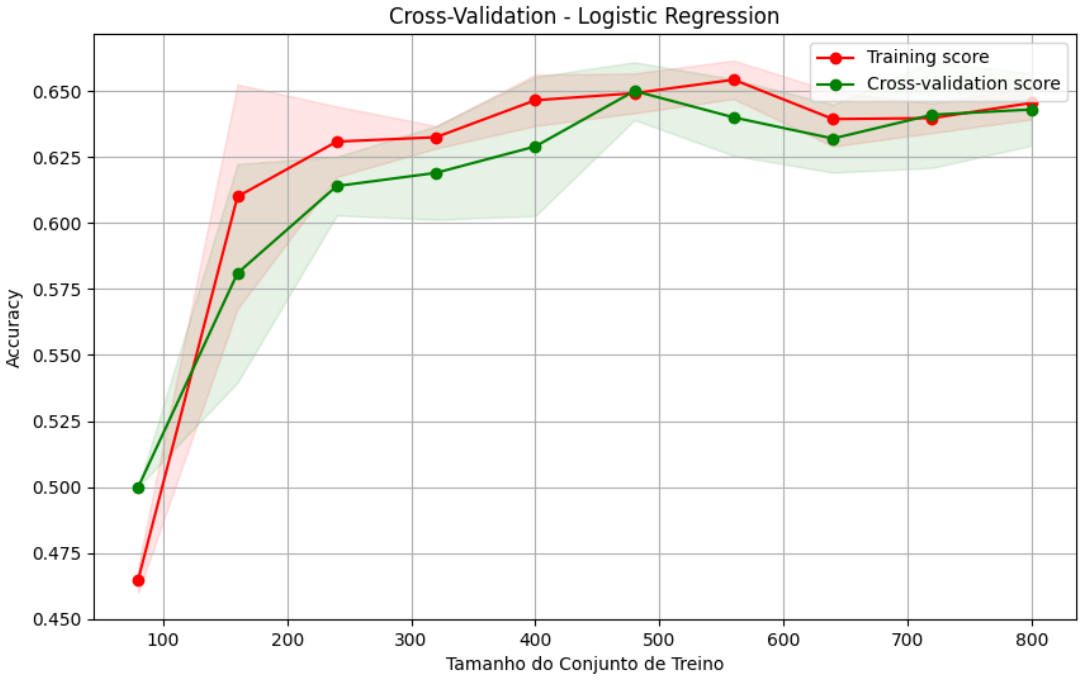

Figura 27 - *Learning Curves – Logistic Regression – Cross-Validation*

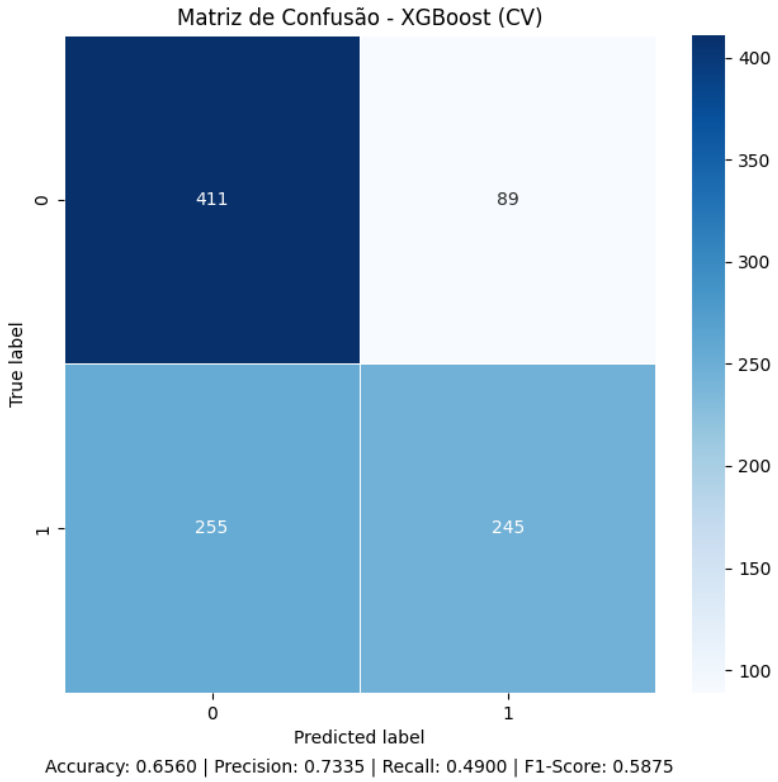

Figura 28 - Matriz de Confusão – *XGBoost – Cross-Validation*

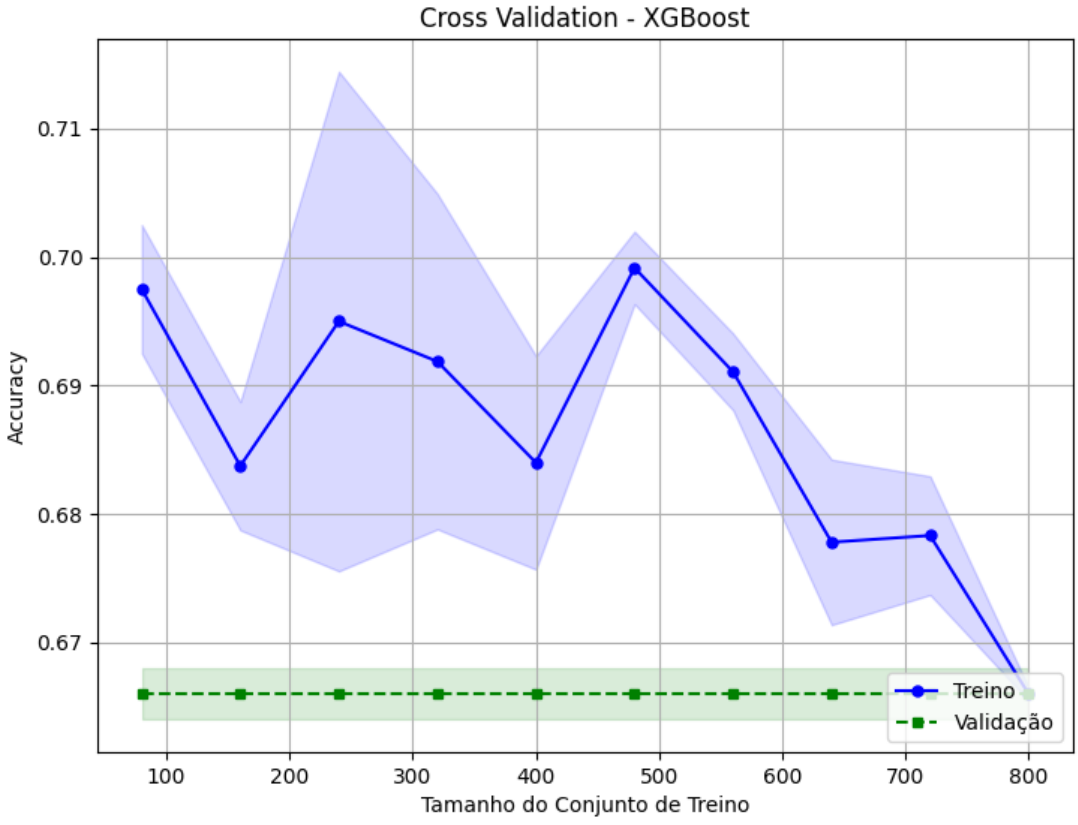

Figura 29 - *Learning Curves – XGBoost – Cross-Validation*

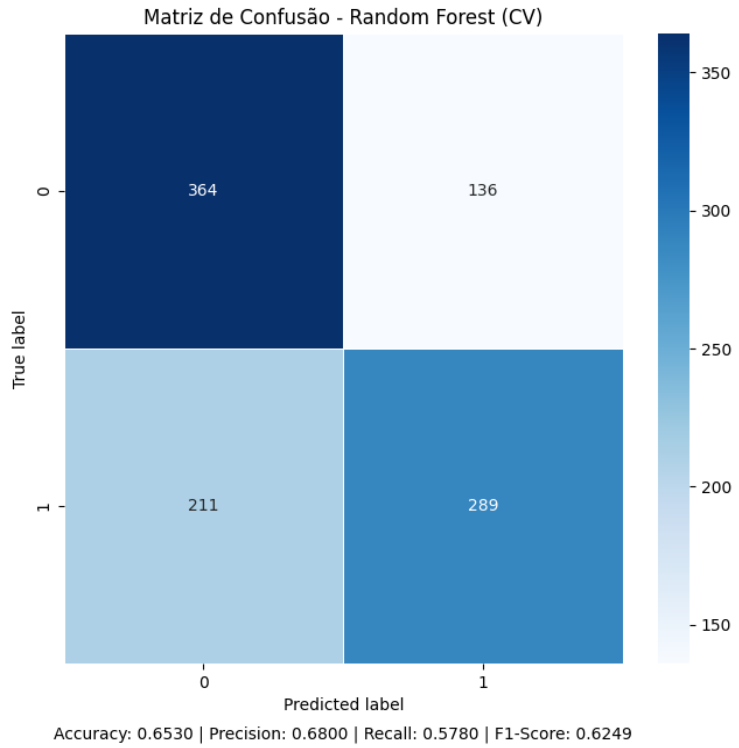

Figura 30 - Matriz de Confusão – *Random Forest – Cross-Validation*

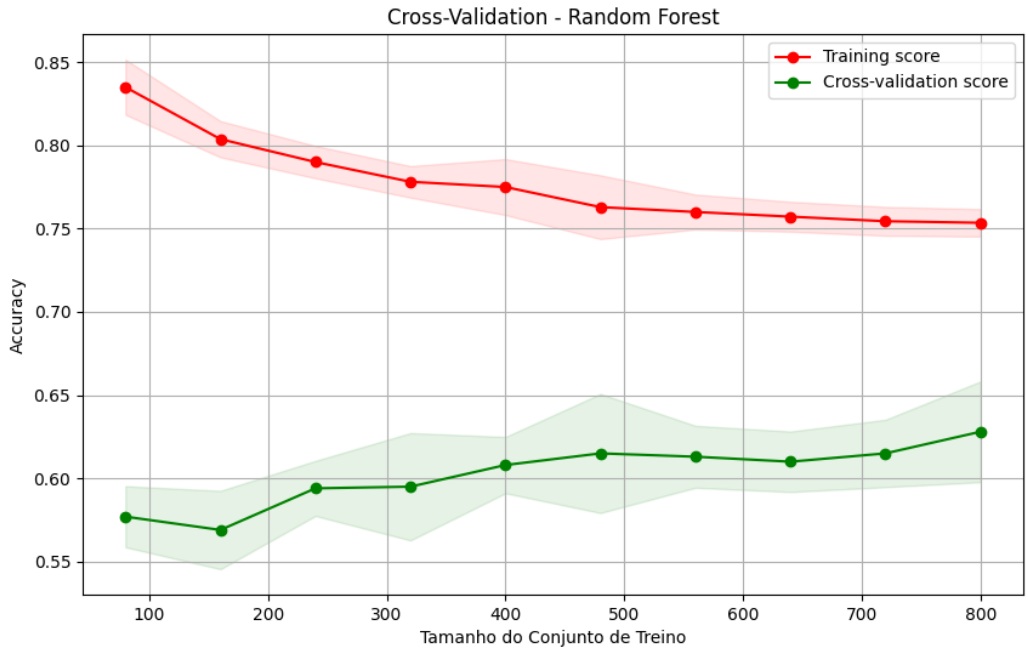

Figura 31 - *Learning Curves – Random Forest – Cross-Validation*

Regressão Logística:

- As *Learning Curves* para a Regressão Logística se aproximam à medida que o tamanho do conjunto de treino aumenta, acontecendo que o treino começa muito mais e vai subindo ao longo, ando que não deve acontecer, tal como a *accuracy* da classificação superar o resultado do treino como se pode verificar, mais ou menos, no Ponto 570 da amostra.
- A matriz de confusão mostra um desempenho semelhante ao do *XGBoost*, com precisão moderadamente alta, mas com espaço para melhorias na sensibilidade (*recall*) do modelo.

XGBoost:

- A *accuracy* de treino varia com oscilações, enquanto a *accuracy* de validação cruzada mostra uma tendência ascendente, o que é positivo, mas as flutuações sugerem que o modelo pode ser sensível à variação dos dados.
- A matriz de confusão mostra uma distribuição mais equilibrada de previsões corretas entre as duas classes.

Random Forest:

- As *Learning Curves* indicam uma *accuracy* alta e estável para o conjunto de treino, mas uma discrepância significativa em relação à *accuracy* da validação cruzada, obtendo *overfitting*, devido a falta de dados para treino
- A matriz de confusão revela que o modelo prevê melhor a classe '0' do que a classe '1', o que pode ser um indício de problemas na classificação.

Durante o contínuo processo de melhoria dos nossos modelos de classificação, iremos nos dedicar a uma análise mais detalhada das ocorrências de falsos negativos e positivos. Tanto esses tipos de erros como suas consequências são capazes de evidenciar lacunas específicas no desempenho do modelo.

Compreender as circunstâncias e características envolvidas nos erros é fundamental para aumentar a precisão do modelo. Investigaremos os padrões recorrentes presentes nas instâncias classificadas erradamente para determinarmos possíveis soluções.

Ao analisar o desempenho dos diversos algoritmos aplicados no nosso modelo de classificação, identificamos um padrão recorrente que suscitou nossa atenção. Observamos que as variáveis referentes ao nível de ensino do pai e da mãe estavam consistentemente presentes entre os falsos positivos e falsos negativos em todos os modelos testados. Esta constatação levou-nos a questionar a relevância e o peso destes atributos na capacidade de classificação do nosso modelo.

Com base nesta análise, decidimos proceder à remoção dos dados relacionados ao ensino do pai e da mãe do nosso conjunto de treino. Esta decisão foi tomada com o objetivo de avaliar se a remoção destas variáveis poderia conduzir a uma melhoria na precisão do modelo.

Após a exclusão destes dados, iremos reexecutar o processo de treino e validação dos algoritmos. Esperamos que esta revisão e simplificação do conjunto de dados possa resultar em um desempenho mais robusto e em métricas de avaliação mais favoráveis.

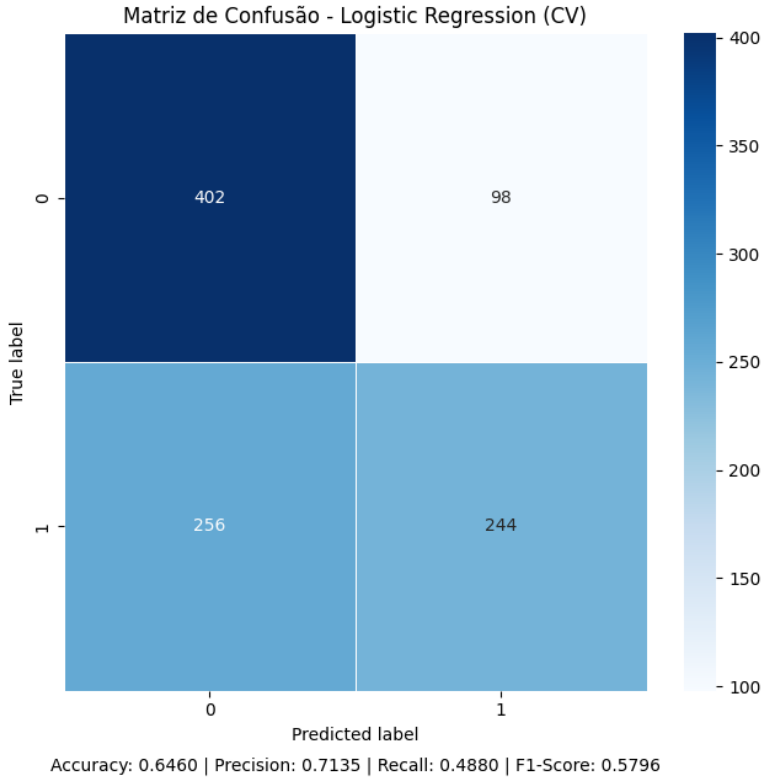

Figura 32 - Matriz de Confusão – *Logistic Regression* – Remover Ensino Pai e Mãe

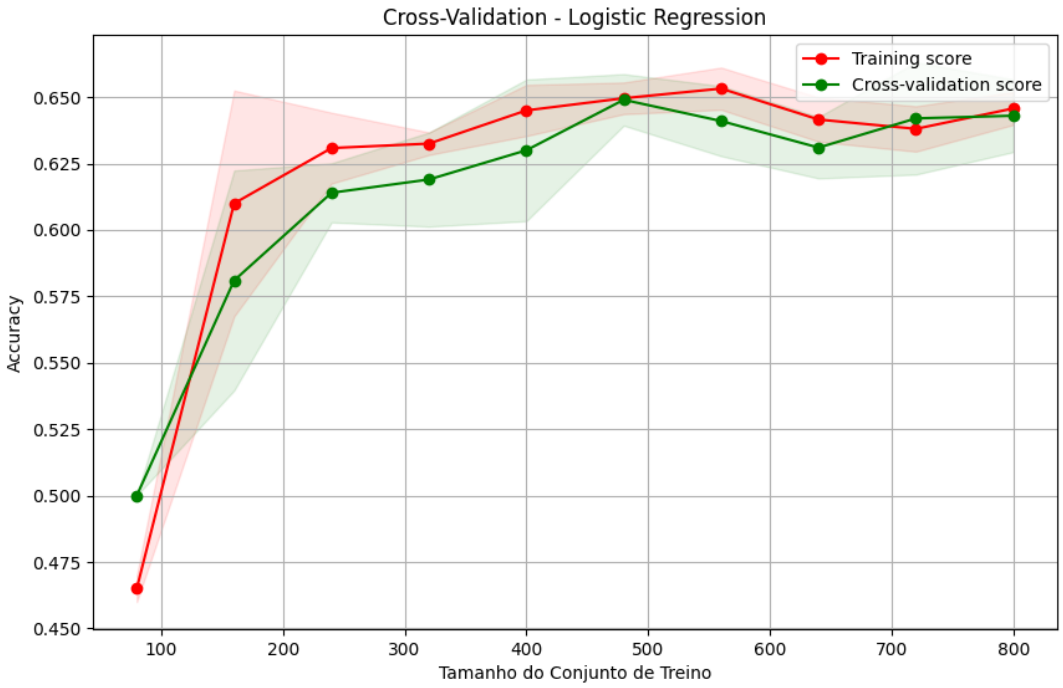

Figura 33 - *Learning Curves – Logistic Regression* – Remover Ensino Pai e Mãe

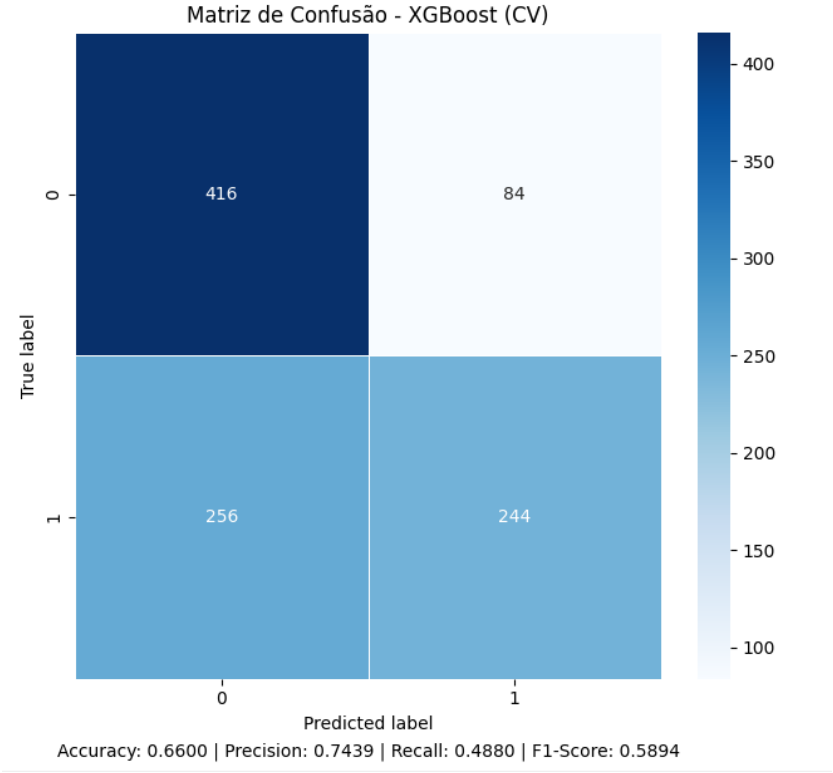

Figura 34 - Matriz de Confusão – *XGBoost* – Remover Ensino Pai e Mãe

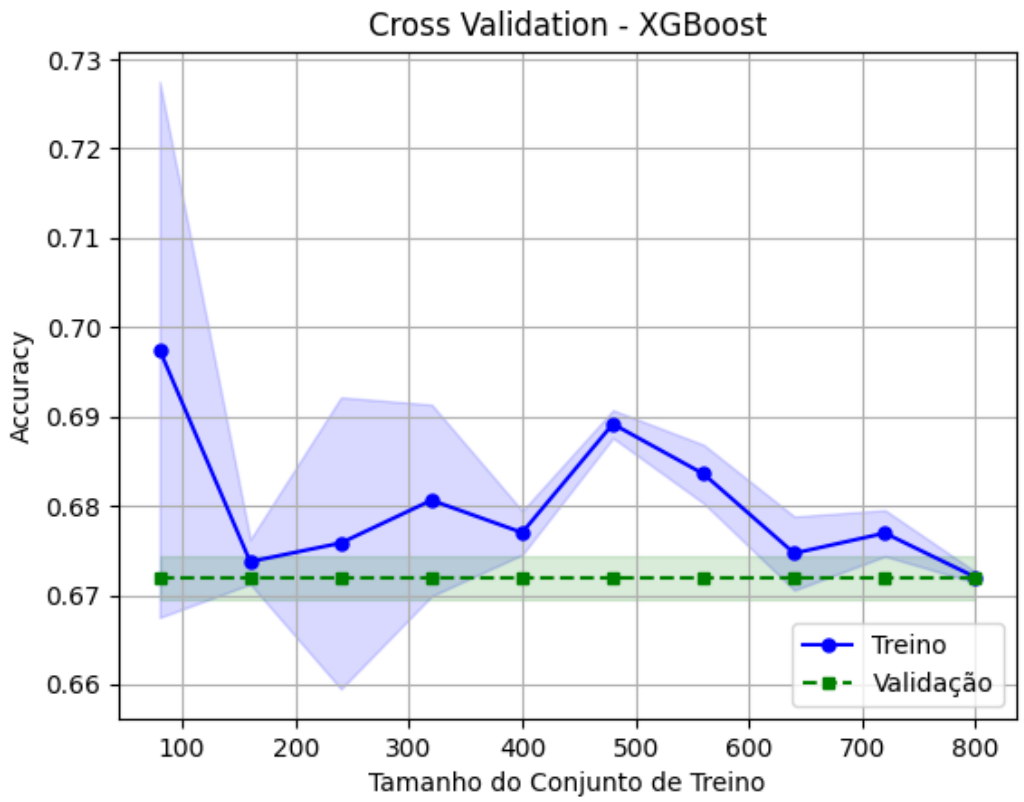

Figura 35 - *Learning Curves – XGBoost* – Remover Ensino Pai e Mãe

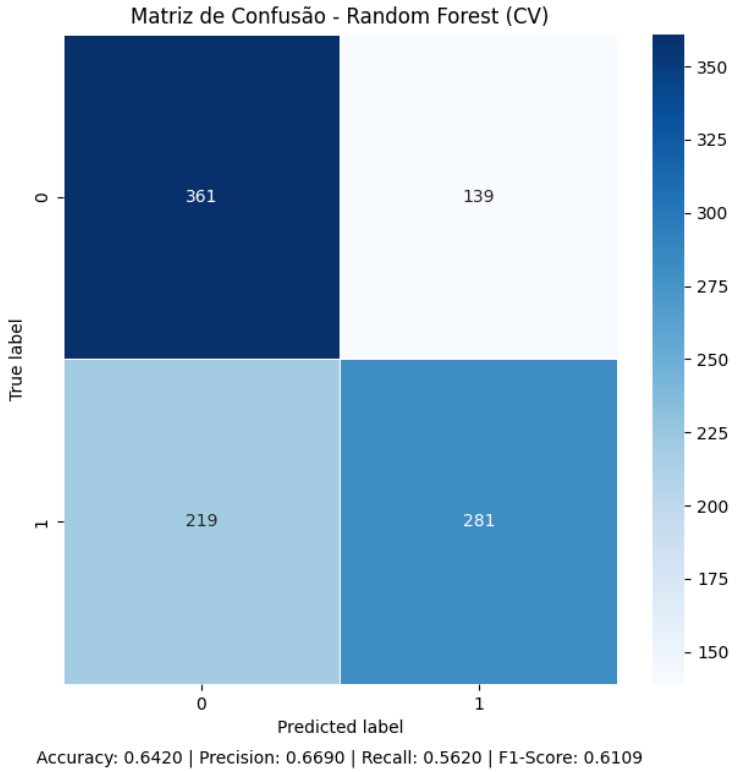

Figura 36 - Matriz de Confusão – *Random Forest* – Remover Ensino Pai e Mãe

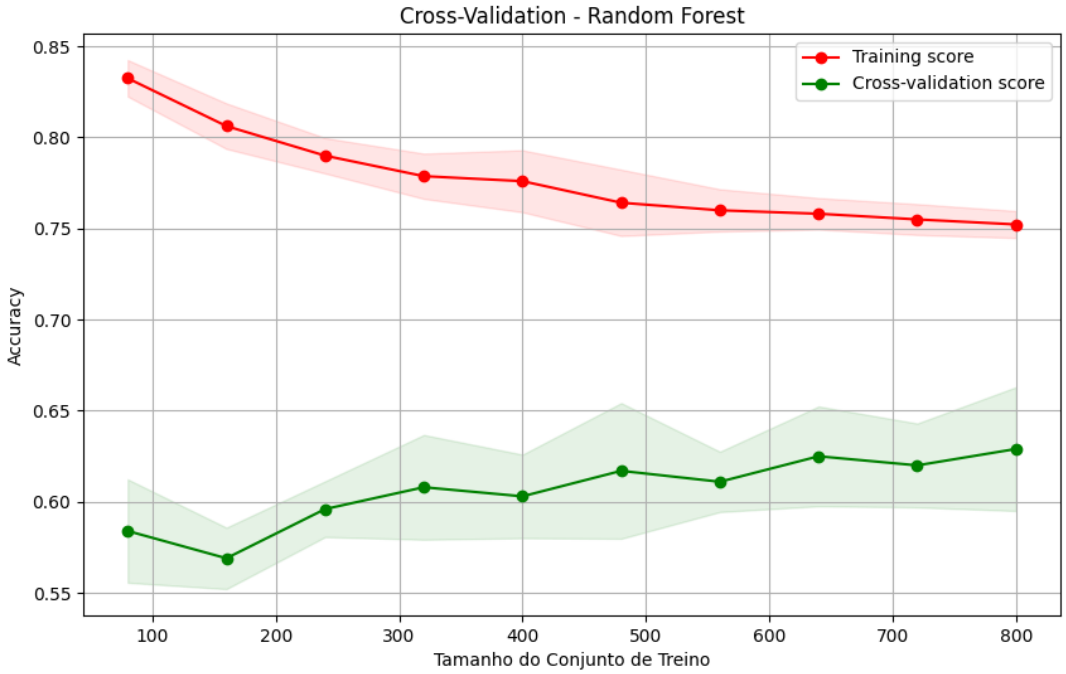

Figura 37 - *Learning Curves – Random Forest* – Remover Ensino Pai e Mãe

Após uma revisão criteriosa do nosso modelo de classificação, a decisão de excluir as variáveis relativas ao nível de ensino dos pais resultou numa melhoria nas métricas de desempenho, embora essa melhoria tenha sido um erro porque foi afetado bastante a classificação da classe 1 'Abandonar', porque o que esta acontecer é que foram retiradas as variáveis que tinham mais correlação com a variável abandonar o ensino superior por isso nos modelos estamos a errar quase a mesma percentagem que a acertar. Por isso esta abordagem que fizemos esta errada e devemos de voltar a colocar estas variáveis inseridas no modelo. Então e ficamos com os mesmos resultados obtidos anteriormente.

5. Avaliação dos Modelos

Enquanto estávamos pesquisando e construindo os nossos modelos sobre o abandono escolar, fizemos uma série testes usando uma variedade de técnicas no campo do *machine learning*. Das alternativas consideradas, foi comprovado que *Random Forest* é a escolha mais promissora em termos dos resultados esperados quando há um aumento significativo no tamanho dos dados utilizados para treinar.

Através da *Learning Curves*, constatou-se que o modelo *Random Forest* possui um desempenho válido quando comparado aos outros algoritmos, demonstrando uma *accuracy* relativamente elevada e estável durante todo o processo. Porém, é importante destacar que a *accuracy* obtida na validação cruzada continua levemente abaixo do esperado, sugerindo assim uma possível melhora no desempenho do modelo com um aumento no volume dos dados utilizados para treino. Essa indicação mostra que o *Random Forest* consegue generalizar bem, porém há espaço para aprimorar suas previsões ao aumentar o tamanho do conjunto de dados usado no treinamento.

O princípio fundamental do Random Forest, segundo o qual "mais dados resultam em melhor desempenho", é particularmente aplicável neste caso. Conforme aumenta-se a quantidade dos exemplos utilizados no treino, espera-se um aumento proporcional na qualidade dos resultados obtidos.

6. Considerações Futuras

Enquanto continuamos a desenvolver nossos modelos visando encontrar características para o abandono do ensino superior, ponderemos sobre as lições extraídas até agora e estabelecemos passos essenciais para o futuro. Constatamos inicialmente a insuficiência na validação da amostra realizada por nós mesmos como consequência dessa falha não nos foi possível identificar os subgrupos relacionados às variáveis assim como suas interações complexas. A existência dessa restrição realçou a importância de uma análise mais minuciosa dos dados disponíveis, para encontrar grupos e padrões ocultos capazes de exercer uma influência significativa nas previsões geradas pelo modelo.

Aprofundamento na Análise de Variáveis:

Estamos focados em conduzir uma análise minuciosa das variáveis existentes no nosso conjunto de dados, e isso é fundamental para as nossas considerações futuras. É necessário compreender tanto as características individuais quanto suas interconexões. No contexto dos modelos complexos que usamos, a interação entre variáveis é uma fonte importante de *insights* e oportunidades para aumentar a precisão do modelo.

Otimização dos Algoritmos de Treino:

Quanto maior for nosso conhecimento em relação aos algoritmos de treino, melhores serão nossos resultados ao ajustarmos os modelos com maior precisão e direcionarmos nosso foco para as técnicas que têm potencial. Essa tarefa envolve explorar novos métodos para seleção e combinação de modelos, bem como realizar ajustes precisos nos Hiper parâmetros que afetam a capacidade do algoritmo em aprender.

Deteção Precoce de *Overfitting*:

Identificar e reduzir o *overfitting* precocemente no processo de treino também é uma área importante que precisa ser aprimorada. Embora seja desejável ter modelos altamente eficientes, é crucial que eles possam verdadeiramente genéricos além dos conjuntos conhecidos e evitar ser meros repositórios de informação de um único conjunto de dados. Trabalhar em métricas e procedimentos de diagnóstico ainda mais refinados para identificar evidências claras do problema do *overfitting*, possibilitando intervenções rápidas e adaptações estratégicas.

7. Conclusão

Este projeto, desenvolvido no contexto da unidade curricular de Inteligência Artificial, revelou-se uma oportunidade valiosa para a aplicação prática e aprofundamento dos conhecimentos teóricos adquiridos ao longo do semestre. Além disso, o desafio de construir e otimizar modelos de máquina permitiu-nos expandir nosso entendimento e habilidades para além do desafio inicial da disciplina.

Durante a execução do trabalho, enfrentamos o desafio de prever o abandono escolar, uma tarefa complexa que exigiu uma análise minuciosa dos dados e uma experimentação rigorosa com diversos algoritmos de *machine learning*. Conseguimos não apenas cumprir os requisitos estabelecidos para o projeto, mas também explorar e implementar técnicas avançadas, incluindo a manipulação de dados, otimização de Hiper parâmetros, entre outros.

Importante destacar que todos os objetivos propostos para o projeto foram alcançados com sucesso. Isto incluiu a realização de testes meticulosos e o uso de ferramentas computacionais apropriadas para validar os nossos modelos. A experiência prática adquirida com a aplicação de diferentes métodos de aprendizagem de máquina, a análise e interpretação dos resultados, bem como o ajuste fino dos modelos, foram fundamentais para uma compreensão mais profunda da área.

Este projeto não apenas reforçou nosso conhecimento teórico em Inteligência Artificial, mas também nos proporcionou *insights* valiosos sobre os desafios reais e as potencialidades do campo de *machine learning*. A capacidade de aplicar teoria à prática em um cenário complexo e dinâmico como o abandono escolar demonstrou a importância e o impacto do aprendizado de máquina na solução de problemas reais e relevantes socialmente.

Em suma, a realização deste projeto foi uma etapa crucial na nossa jornada de aprendizagem, marcando como um ponto significativo de crescimento e desenvolvimento profissional e pessoal na área de Inteligência Artificial.

8. Bibliografia

[1] https://github.com/Zav04/ML.git
<consultado a 23-12-2023>
[2] https://www.geeksforgeeks.org/random-forest-regression-in-python/
<consultado a 23-12-2023>
[3] https://www.geeksforgeeks.org/xgboost/
<consultado a 23-12-2023>
[4] https://www.geeksforgeeks.org/understanding-logistic-regression/
<consultado a 23-12-2023>
[5] https://www.geeksforgeeks.org/how-to-avoid-overfitting-in-machine-learning/
<consultado a 05-01-2024>
[6] https://www.geeksforgeeks.org/using-learning-curves-ml/
<consultado a 12-01-2024>