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/prever 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
- *SMOTE*: *Synthetic Minority Over-sampling Technique*

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. Caracterização dos Dados Utilizados no Modelo de Previsão de Abandono Escolar

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, com 34 atributos distintos. 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.

3.4. Data Augmentation com SMOTE

O *SMOTE* é uma técnica de *oversampling* que gera amostras sintéticas da classe minoritária para combater o problema de desequilíbrio das classes. Ao usar *SMOTE* em um conjunto com algoritmos de classificação, melhora a representação da classe minoritária (alunos que abandonam), permitindo que os modelos aprendam padrões mais genéricos.

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 parecessem 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 treinamento do modelo utilizando o algoritmo de Regressão Logística. A primeira execução foi realizada sem alterações nos Hiper parâmetros padrão.

Estes foram os resultados:

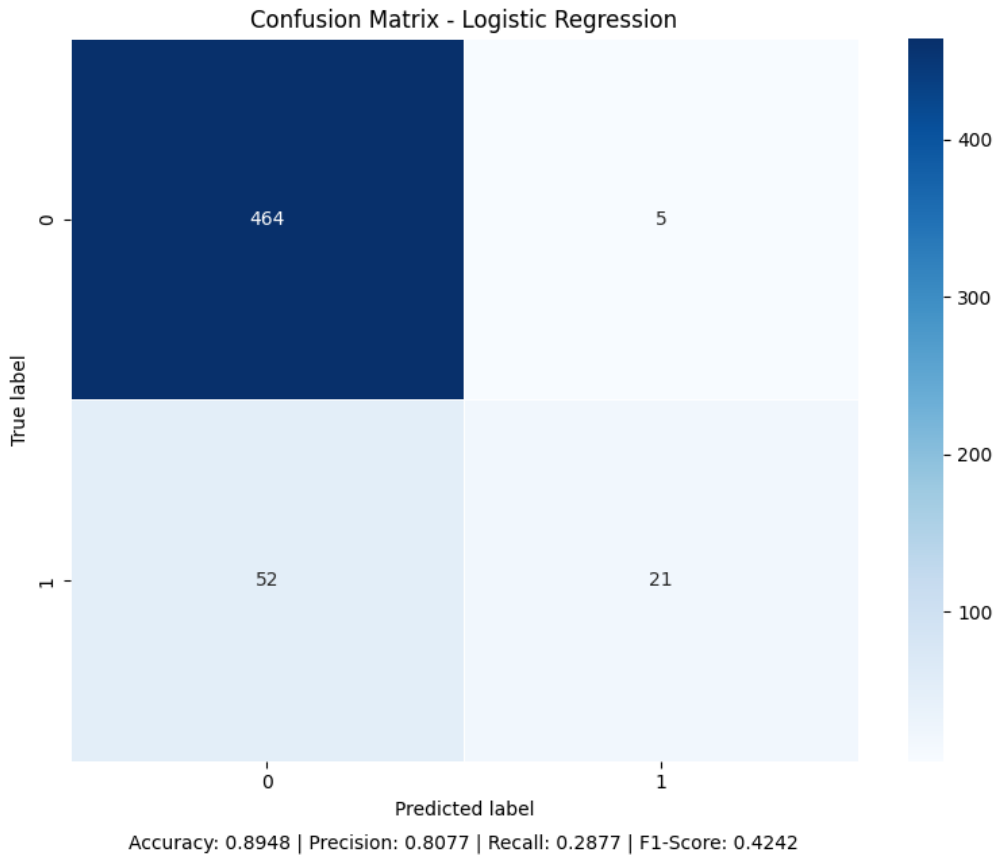

Figura 1 - Matriz de Confusão - Primeiro treino

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 margem para melhorias, uma vez que os indicadores de desempenho ainda estavam distantes do ideal.

Para refinar o desempenho do modelo, iniciamos uma série de experimentos ajustando os Hiper parâmetros de treinamento. Após várias iterações, alcançamos resultados mais otimizados:

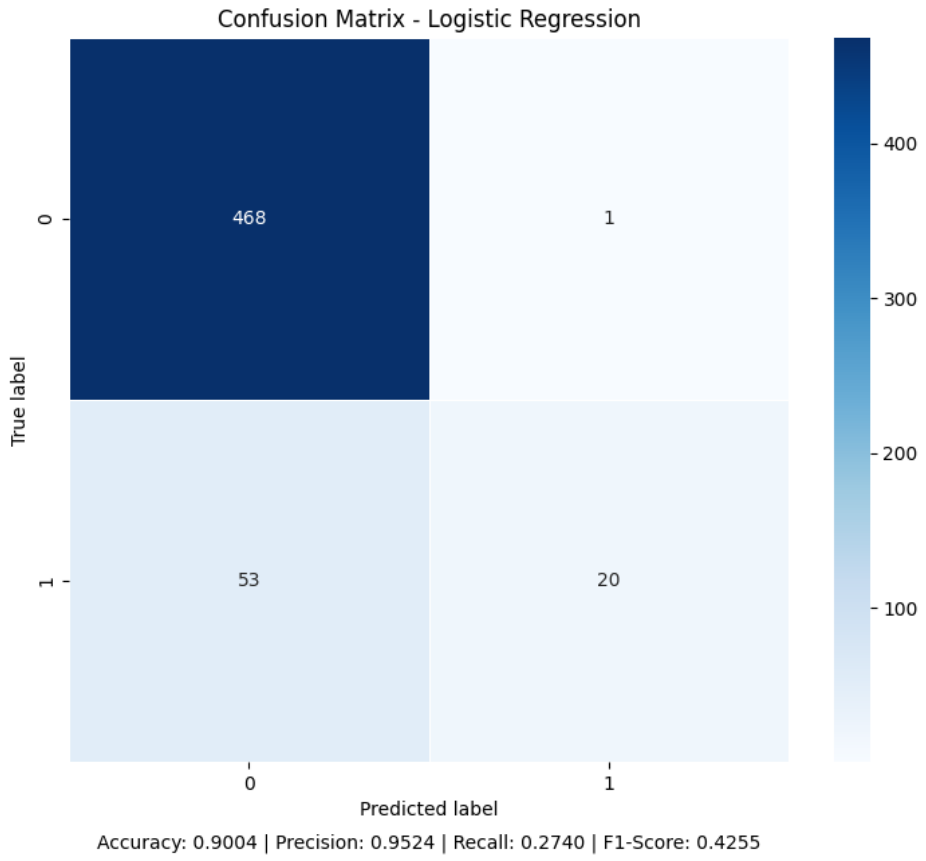

Figura 2 - Matriz de Confusão - Treino com mudança de Hyper Parametros

Esses resultados representam o pico de desempenho que conseguimos extrair da Regressão Logística por meio de ajustes nos Hiper parâmetros.

Entretanto, uma análise do balanço das classes revelou uma distribuição desigual, que poderia estar influenciando a performance do modelo:

Tabela 1 - Balanço entre Classes

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

Diante desse desequilíbrio, optamos pela aplicação de técnicas de *data augmentation*, especificamente o método *SMOTE*, para gerar dados sintéticos e equilibrar as classes. Esperava-se que essa abordagem melhorasse a capacidade do modelo de identificar corretamente os casos de abandono escolar.

Os resultados obtidos foram os seguintes:

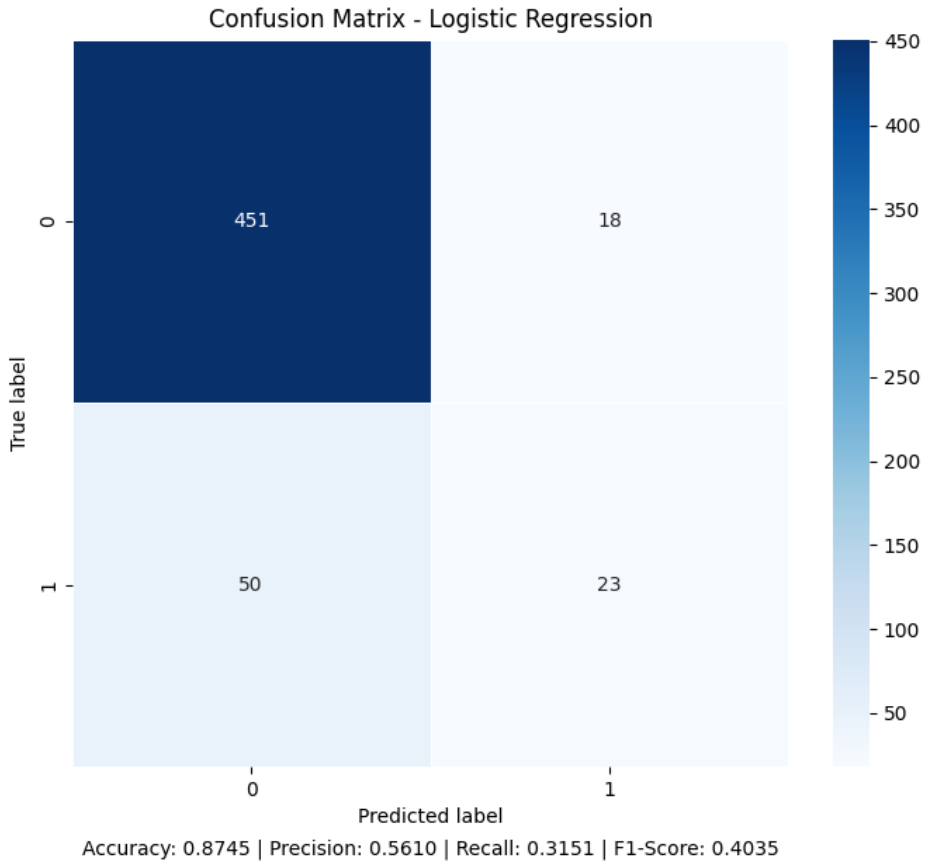

Figura 3 - Matriz de Confusão - Treino com *Data Augmentation*

A introdução de dados sintéticos via técnicas de *data augmentation* não produziu os resultados esperados. As métricas de desempenho do modelo com dados aumentados pelo SMOTE foram inferiores às obtidas com os dados reais. Isso indicou que a adição de dados sintéticos não era benéfica para o nosso caso específico. Com base nessa constatação, decidimos descontinuar a abordagem de *data augmentation*.

Posteriormente, voltamos nossa atenção para a análise e compreensão dos dados existentes. O objetivo era identificar atributos que apresentassem pouca ou nenhuma influência sobre a variável alvo – o abandono escolar – e que, por isso, poderiam ser excluídos para simplificar o modelo e potencialmente melhorar seu desempenho. Utilizamos a

correlação estatística como uma métrica para determinar a força e a direção da relação linear entre cada atributo e a variável alvo.

Tabela 2 - Correlação entre atributos

Atributo	Correlação
Abandono	1.000000
Unnamed	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

A análise da correlação desempenha um papel vital na compreensão do valor informativo de cada atributo. Atributos relacionadas à educação dos pais (*'Ensino Outros Mae/Pai'*) mostram uma correlação moderadamente forte, sugerindo que esses fatores são relevantes para a previsão de abandono.

Curiosamente, variáveis como *'Portugues'*, *'Ensino Basico Pai/Mae'*, que apresentam correlações negativas fortes, podem indicar que alunos com esses perfis têm menor probabilidade de abandonar. Isso poderia refletir, por exemplo, uma maior resiliência ou acesso a recursos de suporte que os ajudam a permanecer na escola.

Por outro lado, atributos com correlação próxima a zero têm uma relação quase insignificante com a variável alvo e podem ser candidatos a exclusão do modelo, simplificando-o e focando apenas nas informações mais impactantes.

A estratégia subsequente se concentrou na remoção de atributos com correlações baixas e na reavaliação do modelo para identificar quaisquer melhorias na performance. A eliminação desses atributos menos correlacionados visa não apenas melhorar a precisão das previsões, mas também tornar o modelo mais interpretável e eficiente computacionalmente.

Os resultados obtidos foram os seguintes:

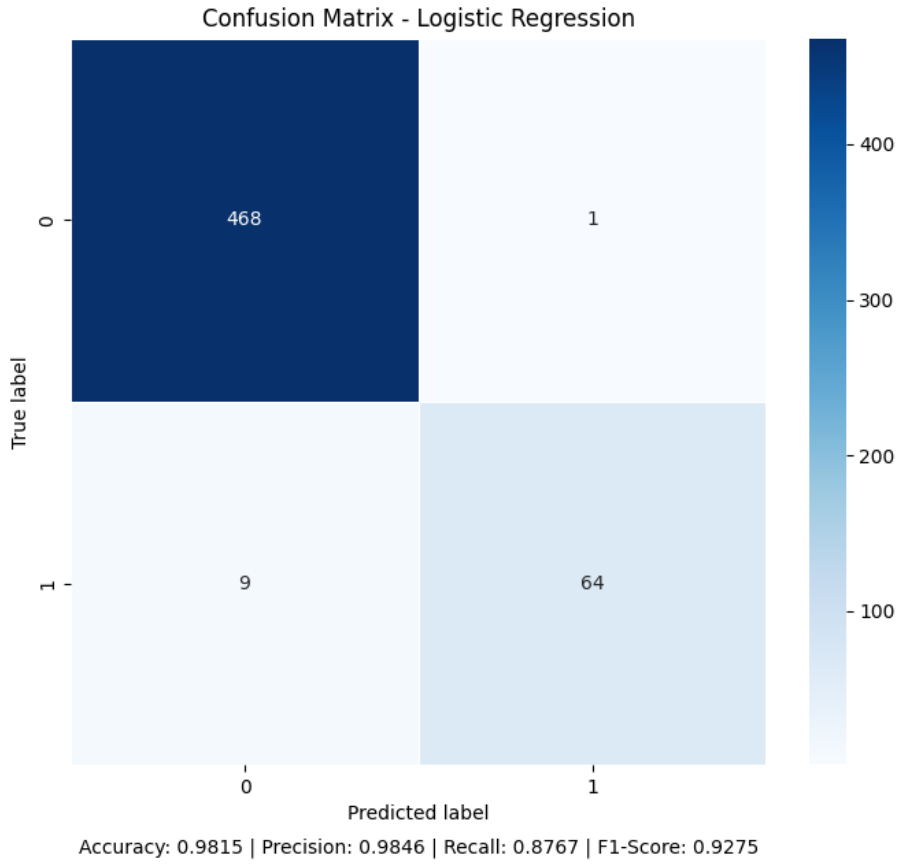

Figura 4 – Matriz de Confusão - Resultados com remoção de variáveis com pouca correlação

Com base na análise de correlação, constatamos que a exclusão de atributos com baixas correlações resultou em uma melhoria significativa do desempenho do modelo. Esta melhoria não só confirmou a hipótese de que atributos menos correlacionados poderiam estar introduzindo ruído nos dados, como também validou a decisão de refinar o conjunto de características para uma maior precisão preditiva.

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.

Com a otimização das características do nosso conjunto de dados, prosseguimos para a fase de treinamento dos diversos algoritmos selecionados para este estudo - Regressão Logística, *XGBoost*, *RandomForest* e Redes Neurais - focando na experimentação e ajuste dos Hiper parâmetros. Este refinamento das variáveis de entrada se mostrou uma estratégia eficaz, servindo como uma fundação consistente para a construção de modelos robustos e confiáveis em todas as abordagens subsequentes de *machine learning* aplicadas ao projeto.

Resultados os restantes modelos:

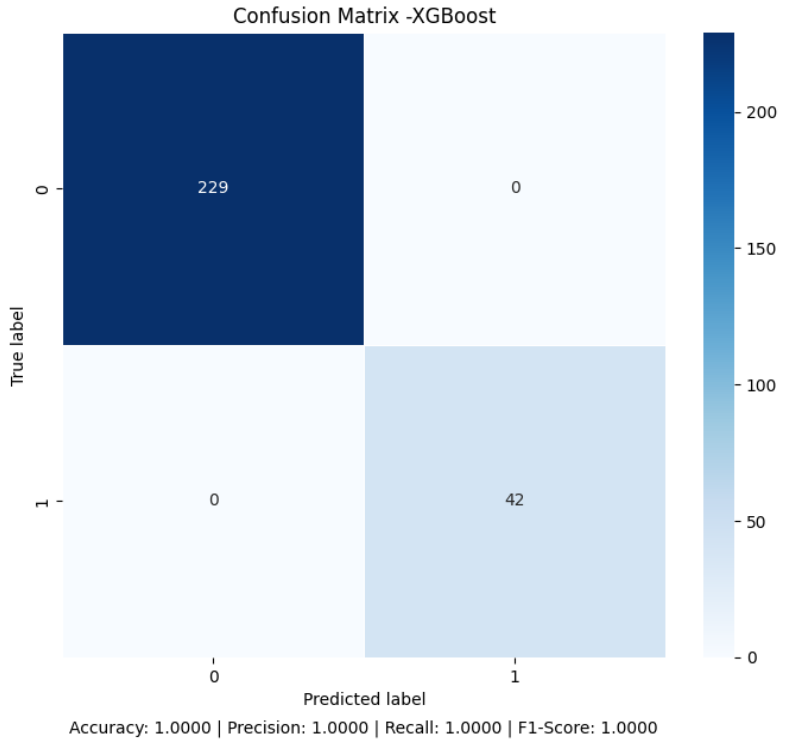

Figura 5 – Matriz de Confusão - XGboost

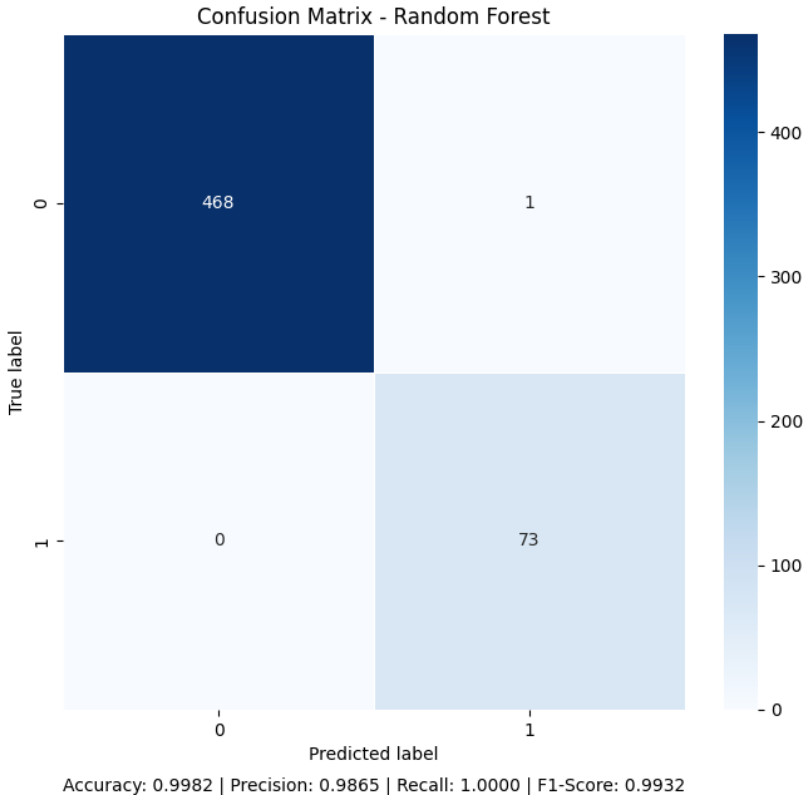

Figura 6 – Matriz de Confusão – Random Forest

No desenvolvimento do modelo de redes neurais, enfrentamos desafios distintos em comparação com os algoritmos tradicionais. A complexidade inerente das redes neurais, decorrente da sua vasta gama de Hiper parâmetros ajustáveis, exigiu uma abordagem mais elaborada e um processo de experimentação intensivo.

As redes neurais oferecem uma flexibilidade significativa: podemos ajustar a arquitetura da rede (número e tamanho das camadas), o método de treino (optimizador e taxa de treino), além de outros parâmetros cruciais como o número de épocas de treino e o tamanho do *batch*. Cada um desses Hiper parâmetros desempenha um papel fundamental no comportamento e na eficácia do modelo, tornando o processo de otimização um desafio complexo e multifacetado.

Realizamos diversas iterações de treino, ajustando cuidadosamente esses Hiper parâmetros na busca por um modelo de rede neural que alcançasse ou superasse a qualidade dos modelos gerados por algoritmos mais tradicionais. Através desse processo rigoroso e detalhado, pudemos explorar profundamente as capacidades e limitações das redes neurais, contribuindo significativamente para o nosso entendimento prático desses sistemas avançados de aprendizado de máquina.

Os melhores resultados que obtivemos:

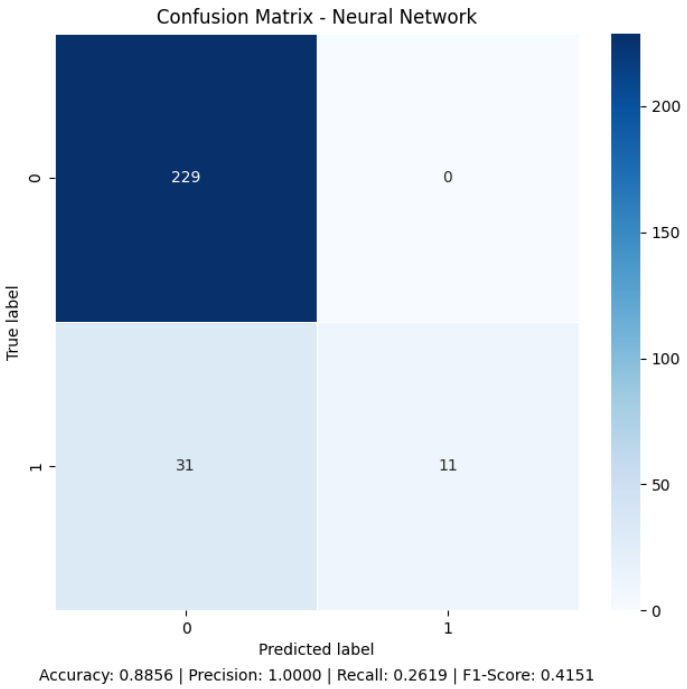

Figura 7 - Matriz de Confusão – Redes Neuronais

5. 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 e a utilização de redes neurais, 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 em nossa jornada de aprendizado, marcando um ponto significativo de crescimento e desenvolvimento profissional na área de Inteligência Artificial e aprendizado de máquina.

6. Bibliografia

[1] https://github.com/Zav04/AI_PROJECT.git
<consultado a 2-12-2023>

[2] https://www.geeksforgeeks.org/breadth-first-search-or-bfs-for-a-graph/
<consultado a 12-10-2023>

[3] https://www.geeksforgeeks.org/a-search-algorithm/
<consultado a 20-10-2023>

[4] https://www.geeksforgeeks.org/greedy-algorithms/
<consultado a 15-01-2023>