



Instituto Federal de Brasília
Campus Estrutural
Licenciatura em Matemática

THIAGO DOS SANTOS FIDELES

**CLASSIFICAÇÃO DE CANDIDATOS NO IFB ATRAVÉS DE LSTM E
MLP USANDO CARTAS DE INTENÇÃO E CURRÍCULOS *LATTES***

Brasília
2023

THIAGO DOS SANTOS FIDELES

**CLASSIFICAÇÃO DE CANDIDATOS NO IFB ATRAVÉS DE LSTM E
MLP USANDO CARTAS DE INTENÇÃO E CURRÍCULOS *LATTES***

Artigo apresentado ao Instituto Federal de Educação, Ciência e Tecnologia de Brasília, *campus* Estrutural, como parte dos requisitos necessários à obtenção do título de Licenciado em Matemática.

Orientador: Me. Pedro Carvalho Brom

Brasília
2023



MINISTÉRIO DA EDUCAÇÃO
Instituto Federal de Educação, Ciência e Tecnologia de Brasília

THIAGO DOS SANTOS FIDELES

**CLASSIFICAÇÃO DE CANDIDATOS NO IFB ATRAVÉS DE LSTM E MLP USANDO
CARTAS DE INTENÇÃO E CURRÍCULOS LATTES**

Trabalho de conclusão de curso apresentado ao Instituto Federal de Educação, Ciência e Tecnologia de Brasília Campus Estrutural como parte dos requisitos necessários à obtenção do título de Licenciado em Matemática.

Aprovado em 07 de dezembro de 2023.

BANCA EXAMINADORA

Prof.(a) Me. Pedro Carvalho Brom

Dr. Rodrigo Figueiredo Abdo

Me. Nilton Cesar de Oliveira Borges

Brasília

2023

Documento assinado eletronicamente por:

- **Pedro Carvalho Brom**, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 18/12/2023 09:50:22.
- **Nilton Cesar de Oliveira Borges**, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 18/12/2023 09:52:57.
- **Rodrigo Figueiredo Abdo**, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 18/12/2023 09:53:06.
- **Antonio Dantas Costa Neto**, COORDENADOR(A) DE CURSO - FUC1 - ES-GRAD-LM, em 18/12/2023 11:51:44.

Este documento foi emitido pelo SUAP em 18/12/2023. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse <https://suap.ifb.edu.br/autenticar-documento/> e forneça os dados abaixo:

Código Verificador: 511756

Código de Autenticação: 052995a64f



Campus Estrutural
Área Especial nº 01, Quadra 16, None, Cidade do
Automóvel/SCIA, ESTRUTURAL / DF, CEP 71.255-200

SUMÁRIO

1	INTRODUÇÃO	6
2	REDES NEURAIIS	8
3	MATERIAL E MÉTODOS	9
3.1	Material	9
3.2	Métodos	10
3.3	Avaliação	11
4	RESULTADOS E DISCUSSÃO	12
4.1	ANÁLISE DO MODELO EXPLICATIVO E PREDITIVO	12
5	CONSIDERAÇÕES FINAIS	13
	REFERÊNCIAS	15

CLASSIFICAÇÃO DE CANDIDATOS NO IFB ATRAVÉS DE LSTM E MLP USANDO CARTAS DE INTENÇÃO E CURRÍCULOS *LATTES*

Thiago dos Santos Fideles¹

Pedro Carvalho Brom²

RESUMO

Este estudo tem como objetivo explorar a eficácia de técnicas avançadas de aprendizado de máquina, particularmente redes neurais recorrentes LSTM (*Long Short-Term Memory*) e MLP (*Multilayer Perceptron*), para prever a classificação na seleção acadêmica de candidatos a programas de especialização baseando-se em suas cartas de intenção e Currículos *Lattes*. Empregando um conjunto de dados de 61 candidatos, o estudo avaliou a capacidade destes modelos em classificar com precisão o potencial acadêmico dos estudantes. Os resultados demonstraram que as rede LSTM apresentam um desempenho moderado, sugerindo uma boa capacidade de identificar os candidatos de sucesso, mas com limitações na discriminação entre os diferentes níveis de desempenho dos candidato. Notavelmente, a rede MLP superou as redes LSTM em todas as métricas avaliadas, indicando seu potencial como um modelo robusto para a previsão de desempenho acadêmico. Através deste estudo, revela-se a aplicabilidade prática de modelos computacionais avançados na seleção acadêmica, sugerindo que tais técnicas podem ser integradas em processos seletivos para prever eficientemente o sucesso dos candidatos. As descobertas apontam para a relevância de continuar a otimização desses modelos e a pesquisa em novas variáveis para aprimorar a precisão da seleção e acompanhamento dos estudantes em ambientes educacionais.

Palavras-chaves: Redes Neurais; Seleção de Estudantes; LSTM; MLP.

ABSTRACT

This study aims to explore the effectiveness of advanced machine learning techniques, particularly Long Short-Term Memory (LSTM) and Multilayer Perceptron (MLP) neural networks, in predicting the academic classification of candidates for specialization programs based on their letters of intent and Curriculum Vitae (*Lattes*). Using a dataset of 61 candidates, the study assessed the ability of these models to accurately classify the academic potential of students. The results demonstrated that LSTM networks exhibited moderate performance, suggesting a good ability to identify successful candidates but with limitations in discriminating between different levels of candidate performance. Notably, the MLP network outperformed LSTM networks in all evaluated metrics, indicating its potential as a robust model for predicting academic performance. Through this study, the practical applicability of advanced computational models in academic selection is revealed, suggesting that such techniques can be integrated into selection processes to efficiently predict candidate success. The findings point to the relevance of continuing the optimization of these models and researching new variables to enhance the accuracy of selection and monitoring of students in educational environments.

¹ Licenciando em Matemática do Instituto Federal de Educação, Ciência e Tecnologia de Brasília (IFB), Brasília/DF - Brasil. E-mail: <thiago.fideles@estudante.ifb.edu.br>.

² Mestre em Estatística pela Universidade de Brasília (UnB). Professor e pesquisador do Instituto Federal de Brasília (IFB), Brasília/DF - Brasil. E-mail: <1521729@ifb.edu.br>.

Keywords: Neural Networks; LSTM; MLP; Student Selection, LSTM, MLP.

1 INTRODUÇÃO

A expansão do ensino superior tem levado as instituições de ensino a enfrentar uma pressão crescente para demonstrar resultados (Casiraghi; Boruchovitch; Almeida, 2022). Um dos aspectos centrais dessa demanda é a capacidade de formar profissionais qualificados que possam responder às necessidades econômicas e tecnológicas atuais (Almeira; Casanova, 2019). No entanto, a evasão e a retenção escolar tornaram-se obstáculos significativos, reduzindo a quantidade de concluintes e, portanto, afetando diretamente a eficácia dessas instituições (Souza, 2021; Silva; Vieira, 2022).

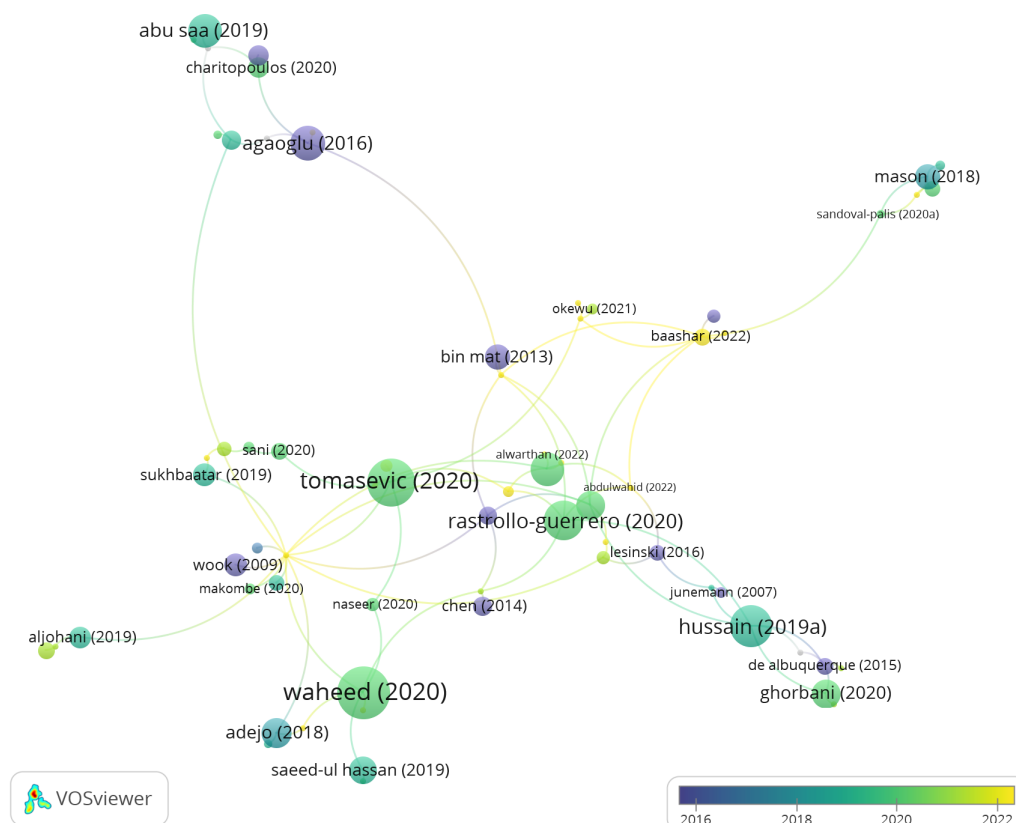
O crescente desenvolvimento das tecnologias de informação (TI) nos últimos anos tem auxiliado no processo de gestão digital dos registros acadêmicos das instituições de ensino superior (IES) e na utilização de estratégias de exploração e análise de grandes quantidades de dados educacionais, também conhecida como mineração de dados educacionais (MDE). Isso permite o acompanhamento, análise e avaliação do processo de aprendizagem do estudante (Souza; Cazella, 2022; Romero; Ventura, 2020). Atualmente, é de responsabilidade dos gestores educacionais e professores a verificação da situação dos discentes e o desenvolvimento de estratégias educacionais para motivá-los (Garcia; Lara; Antunes, 2020; Colpo et al., 2020), além de utilizarem as ferramentas tecnológicas disponíveis para a gestão dos dados acadêmicos e o acompanhamento do desempenho dos alunos.

Esta pesquisa visa a preencher uma lacuna na literatura sobre a eficácia do uso de informações presentes em cartas de intenção e Currículos *Lattes* na predição do desempenho acadêmico dos alunos. Para isso, foi conduzido um estudo bibliométrico, fundamentado na teoria do enfoque meta analítico consolidado (Mariano; Santos, 2017), que envolve a preparação da pesquisa, apresentação e interrelação dos dados, detalhamento, modelo integrador e validação por evidências. Este estudo utilizou um conjunto de termos relacionados à predição do rendimento acadêmico na base de dados *Web of Science*, que identificou 38 estudos publicados nos últimos 5 anos. Essas informações foram processadas através do software VOSviewer versão 1.6 para construção de redes bibliométricas e visualização de dados.

No estudo realizado por Waheed (2020), o objetivo principal foi avaliar a eficácia dos Ambientes Virtuais de Aprendizagem (AVAs) na previsão do desempenho dos alunos, visando a possibilitar intervenções oportunas por parte dos instrutores e fornecer um suporte pedagógico adequado. Por meio da utilização de um modelo de rede neural profunda, os resultados revelaram uma precisão entre 84% e 93% na classificação do desempenho dos estudantes. Para a análise, foram considerados como conjunto de dados o desempenho nas avaliações, os padrões de cliques registrados na plataforma online de aprendizagem, bem como informações demográficas.

Rastrollo-Guerrero, Gomez-Pulido e Duran-Dominguez (2020) conduziram uma revisão de 64 artigos que abordam a previsão do comportamento do estudante no ambiente acadêmico. Desses, 70% se concentram na previsão do desempenho de estudantes universitários. Foi observado que a inclusão de dados demográficos e notas em avaliações é considerada um conjunto de dados de entrada robusto e relevante. Obteve-se, como resultado, que a técnica mais utilizada para prever o comportamento dos estudantes é a aprendizagem supervisionada, com destaque para o algoritmo *Support Vector Machines* (SVM). Por outro lado, as redes neurais são menos utilizadas, apesar de apresentarem alta precisão.

Figura 1 – Estudo bibliométrico - Análise de citações



Fonte: Elaborado pelo autor

No trabalho de Tomasevic, Gvozdenovic e Vranes (2020), acerca da aplicação de técnicas de aprendizado de máquina supervisionado na previsão do desempenho dos alunos, foi constatado que a utilização de redes neurais profundas, em um conjunto os dados de engajamento do aluno e histórico de desempenho, apresentou resultados promissores. Por outro lado, verificou-se que o uso de dados demográficos não demonstrou uma influência significativa na precisão das previsões.

Com base na análise prévia da literatura e na necessidade de melhorar a previsão do desempenho acadêmico, propomos investigar os elementos específicos presentes nas cartas de intenção (por exemplo, experiência prévia na área, habilidades de pesquisa demonstradas, clareza de objetivos acadêmicos) e nos Currículos *Lattes* (por exemplo, publicações anteriores, projetos de pesquisa, experiências de ensino) como potenciais indicadores para prever com precisão o desempenho acadêmico dos alunos. Neste contexto, vislumbramos a possibilidade não apenas de reconhecer a existência destes elementos, mas também de identificar padrões que, uma vez compreendidos, possam ser utilizados para antecipar, com um grau substancial de acurácia, o desempenho futuro dos estudantes.

A carta de intenção é um documento solicitado aos candidatos nos processos seletivos de programas de pós-graduação, com o objetivo de compreender a motivação de ingresso no curso, as experiências acadêmicas e profissionais mais relevantes, metas de carreira e como isso se relaciona com o curso desejado (Dantas, 2022). Já o Currículo *Lattes* é a maior plataforma de registro de vida pregressa e atual de estudantes e pesquisadores do Brasil (Brito;

Quoniam; Mena-Chalco, 2016), sendo adotado hoje pela maioria dos IES, além de ser um dos pré requisitos para obtenção de bolsas e auxílios públicos em cursos de pós-graduação (Nobre; Freitas, 2017; Lopes, 2022). Neste estudo, portanto, exploraremos a possibilidade de prever a classificação dos estudantes com base em informações contidas nas cartas de intenção e nos Currículos *Lattes* e verificar se essas informações podem ser utilizadas para prever o rendimento acadêmico dos estudantes, utilizando um modelo de redes neurais profundas.

2 REDES NEURAIS

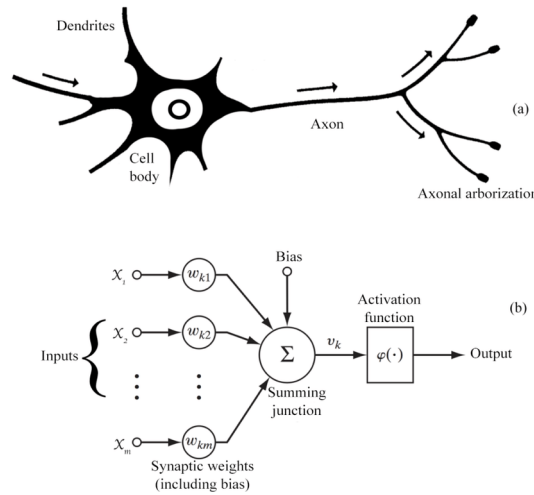
As redes neurais, uma classe de modelos computacionais no campo de *Machine Learning* (ML), tiveram origem na pesquisa em inteligência artificial (IA) com o objetivo de simular o funcionamento do cérebro humano. Em 1943, McCulloch e Pitts desenvolveram o primeiro modelo matemático de neurônio, seguido pelo *perceptron* de Frank Rosenblatt em 1958, considerado o precursor das redes neurais modernas (Emmert-Streib et al., 2020; Schmidhuber, 2015; Alom et al., 2019). Apesar dos desafios iniciais, como dificuldade de treinamento e a limitação computacional em problemas mais complexos (Emmert-Streib et al., 2020; Schmidhuber, 2015; Alom et al., 2019), o ressurgimento das redes neurais ocorreu na década de 1980 com a descoberta do algoritmo de retropropagação (Emmert-Streib et al., 2020; Schmidhuber, 2015; Alom et al., 2019). Isso possibilitou o treinamento eficiente de redes neurais de múltiplas camadas, impulsionando sua popularidade graças ao aumento do poder computacional e conjuntos de dados em expansão.

Os neurônios artificiais são componentes fundamentais no desenvolvimento de redes neurais, baseados nos neurônios biológicos, são compostos pelas entradas ponderadas, função de ativação e saída (Emmert-Streib et al., 2020; Schmidhuber, 2015; Alom et al., 2019; Munir et al., 2019). Cada entrada é multiplicada por um peso que reflete sua relevância para o neurônio. A soma ponderada é então processada pela função de ativação, introduzindo não-linearidade na resposta. As redes neurais são compostas por camadas interconectadas de neurônios, em que a camada de entrada recebe dados brutos e os transmite para a camada oculta, responsável pelo processamento intermediário, até a camada de saída que gera a resposta final. Os pesos das conexões são ajustados durante o treinamento para otimizar o desempenho.

As redes neurais têm ampla aplicação em áreas como processamento de imagens e visão computacional (Munir et al., 2019), reconhecimento de padrões (Ledig et al., 2017), processamento de linguagem natural (Sutskever; Vinyals; Le, 2014), análise e mineração de dados (Waheed et al, 2020; Rastrollo-Guerrero; Gomez-Pulido; Duran-Dominguez, 2020). As células *Long Short-Term Memory* (LSTM) são um tipo de rede neural recorrente (RNN) desenvolvidas para lidar com dependências de longo prazo em sequências (Yong et al., 2019; Munir et al., 2019; Sagheer; Mostafa, 2019).

As células LSTM possuem componentes-chave, como a porta de entrada (*input gate*), a porta de esquecimento (*forget gate*), a porta de saída (*output gat*) e o estado da célula (*cell state*) (Sagheer; Mostafa, 2019). Esses componentes trabalham em conjunto para controlar a entrada de informações, o esquecimento de estados anteriores e a produção de saídas relevantes. Após o cálculo das portas, o estado da célula é atualizado considerando o estado anterior, a entrada atual e a saída da porta de esquecimento. A saída final da célula LSTM é determinada pela porta de saída e pode ser utilizada para fazer previsões sobre o próximo elemento em uma sequência (Yong et al., 2019; Munir et al. 2019). As células LSTM têm se mostrado extremamente eficazes em diversas aplicações, incluindo processamento de linguagem natural (Xu et al., 2019; Jang et al., 2020), reconhecimento de fala e tradução automática (Orugh; Viriri; Adegun, 2022; Su

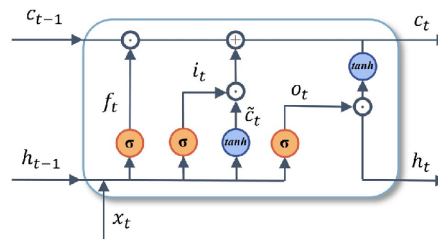
Figura 2 – Similaridade entre redes neurais biológicas e artificiais



Fonte: (Akgun; Demir, 2018)

et al., 2020), regressão e classificação (Ge; Lin; Yu, 2019; Liu; Guo, 2019; Chang; Masterson, 2020; Zhang et al., 2019).

Figura 3 – Arquitetura da célula LSTM



Fonte: (Ji et al., 2023)

3 MATERIAL E MÉTODOS

Este capítulo apresenta os procedimentos metodológicos desta pesquisa com o intuito de atingir os objetivos propostos. A seguir estão descritas as etapas, os microdados utilizados e as ferramentas necessárias.

3.1 Material

Este estudo utiliza um conjunto de dados inicial consistindo em 61 observações, coletados pela coordenação do curso de Educação, Matemática e Tecnologias do Instituto Federal de Educação, Ciência e Tecnologia de Brasília, *Campus* Estrutural. Estes dados, incluindo informações sobre os candidatos ao curso de pós-graduação oferecido no Campus, como o ID do candidato, instituição de origem, tipo de vaga para a qual se candidataram, a carta de intenção e o Currículo *Lattes*, foram cedidos para o propósito deste estudo em maio de 2023. A utilização desses dados está estritamente em conformidade com a Lei Geral de Proteção de Dados (LGPD) (BRASIL, 2018), conforme o artigo 11º, inciso II, alínea “c”, que autoriza o trata-

mento de dados pessoais sensíveis para fins de estudos por órgãos de pesquisa, com a condição de anonimização dos dados sempre que possível.

3.2 Métodos

Este estudo focou no desenvolvimento e análise de duas redes neurais distintas. A primeira rede neural, baseada no algoritmo LSTM, é composta por sete camadas. A camada inicial é uma camada de incorporação que processa palavras provenientes do processo de tokenização. As camadas seguintes, especificamente a segunda e a quarta, consistem em 32 unidades de LSTM cada. As terceira e quinta camadas são de *Dropout*, uma técnica de regularização para redes neurais, ajustadas em 0.5. A sexta camada é composta por 32 unidades de LSTM bidirecional, e a sétima e última camada apresenta duas unidades de saída com ativação do tipo “*softmax*”.

A segunda rede neural, implementada com o modelo *Multilayer Perceptron* (MLP), é formada por cinco camadas. A primeira camada, de entrada, recebe a quantidade de parâmetros da base de dados. As segunda e quarta camadas são constituídas por 32 unidades com ativação “ReLU”, enquanto a terceira camada possui 64 unidades também com ativação “ReLU”. A quinta camada, a camada de saída, tem uma única unidade com ativação *sigmoid*. Para a definição dos hiperparâmetros arquiteturais de ambas as redes, recorreu-se à ferramenta AutoML¹. Em relação à função de perda adotada, utilizou-se *binary crossentropy*,

$$Binary\ Crossentropy = -\frac{1}{N} \sum_{i=1}^N [y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)], \quad (3.1)$$

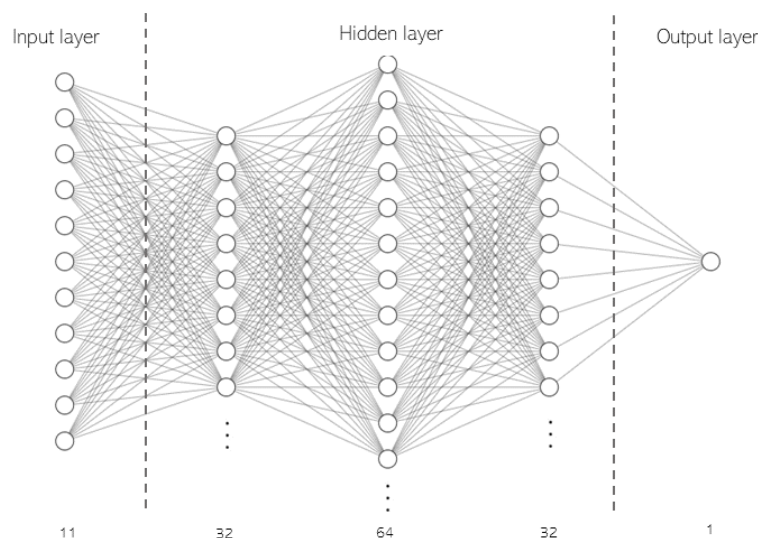
onde N é o número total de observações no conjunto de dados, y_i é a etiqueta verdadeira da i -ésima amostra rotuladas por 0 (não classificado) e 1 (classificado), $\hat{y}_i$ é a probabilidade prevista pelo modelo de que a i -ésima amostra pertence à classe 1 e $\log$ é o logaritmo natural.

Essa função é amplamente utilizada em contextos de classificação binária e seu principal objetivo é avaliar o desempenho do modelo na predição da probabilidade associada a cada uma das classes. Essencialmente, ela impõe penalidades às previsões que divergem das classificações verdadeiras. Em relação à otimização, optou-se pelo emprego do otimizador *Adaptive Moment Estimation* (ADAM)². Este representa um método eficaz para a atualização dos pesos das redes neurais, fundamentando-se em estimativas adaptativas dos momentos de primeira e segunda ordem, o que contribui para aprimorar a precisão do modelo.

¹ O AutoML é um método automatizado para a seleção e otimização de modelos e parâmetros de aprendizado de máquina, facilitando o processo de desenvolvimento e garantindo uma maior eficiência.

² Algoritmo de otimização em redes neurais que combina técnicas de *momentum*, que ajuda a acelerar o algoritmo de gradiente descendente na direção correta e suavizar as oscilações, e *Root Mean Square Propagation* (RMSprop), que ajusta a taxa de aprendizado de forma adaptativa para cada parâmetro, dando mais importância para os parâmetros mais informativos, para atualizar os pesos da rede de forma eficiente, acelerando a aprendizagem e ajustando taxas de aprendizado individualmente para cada parâmetro.

Figura 4 – Arquitetura da Rede Neural MLP



Fonte: Elaborado pelo autor

3.3 Avaliação

Os resultados são avaliados com base em duas métricas amplamente reconhecidas em tarefas que envolvem problemas de classificação. A matriz de confusão é uma ferramenta crucial na avaliação de modelos de classificação em aprendizado de máquina. Ela apresenta o desempenho do modelo de forma tabular, ajudando a entender como ele classifica as instâncias em diferentes categorias. A tabela é dividida em quatro partes principais:

1. Verdadeiro Positivo (VP): Estes são os casos onde o modelo prevê corretamente a presença da classe de interesse (positiva).
2. Falso Negativo (FN): Ocorre quando o modelo prevê incorretamente a ausência da classe de interesse, mas na realidade, ela está presente.
3. Falso Positivo (FP): Acontece quando o modelo prevê erroneamente a presença da classe de interesse, enquanto na realidade, ela está ausente.
4. Verdadeiro Negativo (VN): São os casos em que o modelo acerta ao prever a ausência da classe de interesse.

Tabela 1 – Matriz de Confusão

	Predito	
	Positivo	Negativo
Real Positivo	VP	FN
Negativo	FP	VN

Fonte: Elaborado pelo autor

A matriz é uma forma direta de visualizar a frequência de acertos e erros do modelo, permitindo avaliar métricas importantes como *precision*, *recall* e a taxa de acerto. É especial-

mente útil em cenários onde as classes estão desbalanceadas ou quando os custos de diferentes tipos de erros variam significativamente.

A primeira métrica é a Acurácia (ACC),

$$ACC = \frac{VP + VN}{VP + VN + FP + FN}, \quad (3.2)$$

onde VP representa o total de verdadeiros positivos, VN representa o total de verdadeiros negativos, FP são os falsos positivos e FN os falsos negativos. É usada para comparar os valores reais de uma instância com os valores previstos pelo algoritmo. A segunda métrica avaliada é a F1-Score (F1),

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (3.3)$$

que representa a média harmônica entre as métricas de *Precision* e *Recall*. Tanto a acurácia quanto a F1 são expressas em uma escala de 0 a 1, na qual um valor maior indica um desempenho melhor. A acurácia em geral é utilizada quando temos dados balanceados e o F1 é usado como avaliação nos casos de dados desbalanceados.

Finalmente, a implementação de ambos os modelos foi realizada utilizando a linguagem de programação Python e a biblioteca Keras, uma escolha adequada devido à sua flexibilidade e capacidade de facilitar a experimentação com diferentes arquiteturas de redes neurais.

4 RESULTADOS E DISCUSSÃO

4.1 ANÁLISE DO MODELO EXPLICATIVO E PREDITIVO

O presente estudo foi concebido com o objetivo de avaliar a capacidade das cartas de intenção e das informações contidas nos Currículos *Lattes* em prever o rendimento acadêmico de estudantes nos cursos de especialização em Educação, Matemática e Tecnologias oferecidos pelo Instituto Federal de Educação, Ciência e Tecnologia de Brasília, *campus* Estrutural. Para tanto, empregou-se modelos avançados de aprendizado de máquina, nomeadamente redes neurais LSTM e MLP.

Tabela 2 – Resultado das métricas de avaliação

Métricas	LSTM - Currículo <i>Lattes</i>	LSTM - Carta de intenção	MLP
Acurácia	0.5455	0.5455	0.9091
Precisão	0.5455	0.5455	0.8750
<i>Recall</i>	1.0000	1.0000	1.0000
F1-Score	0.7059	0.7059	0.9333
ROC AUC	0.7667	0.6667	0.8393

Fonte: Elaborado pelo autor.

Com um conjunto de dados inicial composto por 61 amostras de candidatos, a pesquisa visou não apenas identificar elementos-chave nas cartas de intenção e Currículos *Lattes* relacionados ao rendimento acadêmico e ao processo de seleção e admissão, mas também implementar modelos baseados em LSTM e MLP para prever o rendimento e a classificação dos estudantes, respectivamente.

A análise realizada com a rede LSTM para as cartas de intenção resultou em uma Acurácia e Precisão de 0.5455, com um *Recall* de 1.0000, indicando que o modelo foi capaz de

identificar todos os candidatos que seriam bem-sucedidos, mas com uma capacidade limitada de precisão na classificação. O F1-score obtido foi de 0.7059 e a área sob a curva ROC (AUC) foi de 0.7667, o que sugere uma capacidade razoável do modelo em discriminar entre candidatos de sucesso e outros.

Por sua vez, a rede LSTM aplicada aos Currículos *Lattes* também alcançou uma Acurácia e Precisão de 0.5455, com um *Recall* perfeito de 1.0000 e um F1-score idêntico de 0.7059. A AUC para esta análise foi de 0.6667, ligeiramente inferior àquela obtida para as cartas de intenção, o que pode indicar uma necessidade de refinamento na seleção de características ou na parametrização do modelo para essa categoria de dados.

Notavelmente, a rede MLP demonstrou superioridade substancial em relação as redes LSTM, com uma acurácia de 0.9091, precisão de 0.8750, *Recall* de 1.0000, F1-score de 0.9333 e uma AUC de 0.8393. Estes resultados ressaltam a robustez do modelo MLP na previsão da classificação e seleção de estudantes, fornecendo um instrumento com alto grau de confiabilidade para uso em processos seletivos acadêmicos.

Esses resultados ressaltam o MLP como um modelo de destaque na previsão de desempenho acadêmico. A consistência de alto *Recall* em ambas as abordagens LSTM e MLP implica que os modelos são altamente eficazes na identificação de todos os candidatos qualificados, embora a precisão e a AUC variem, indicando diferenças na capacidade de discriminação entre os candidatos de sucesso e os demais. A superioridade do modelo MLP em todas as métricas avaliadas sinaliza sua aplicabilidade promissora na seleção e classificação de estudantes para programas de especialização. A alta acurácia e F1-score sugerem que o modelo é capaz de balancear bem a precisão e a sensibilidade, minimizando tanto os falsos positivos quanto os falsos negativos, o que é crucial em contextos acadêmicos onde a identificação precisa do potencial de sucesso do aluno é essencial.

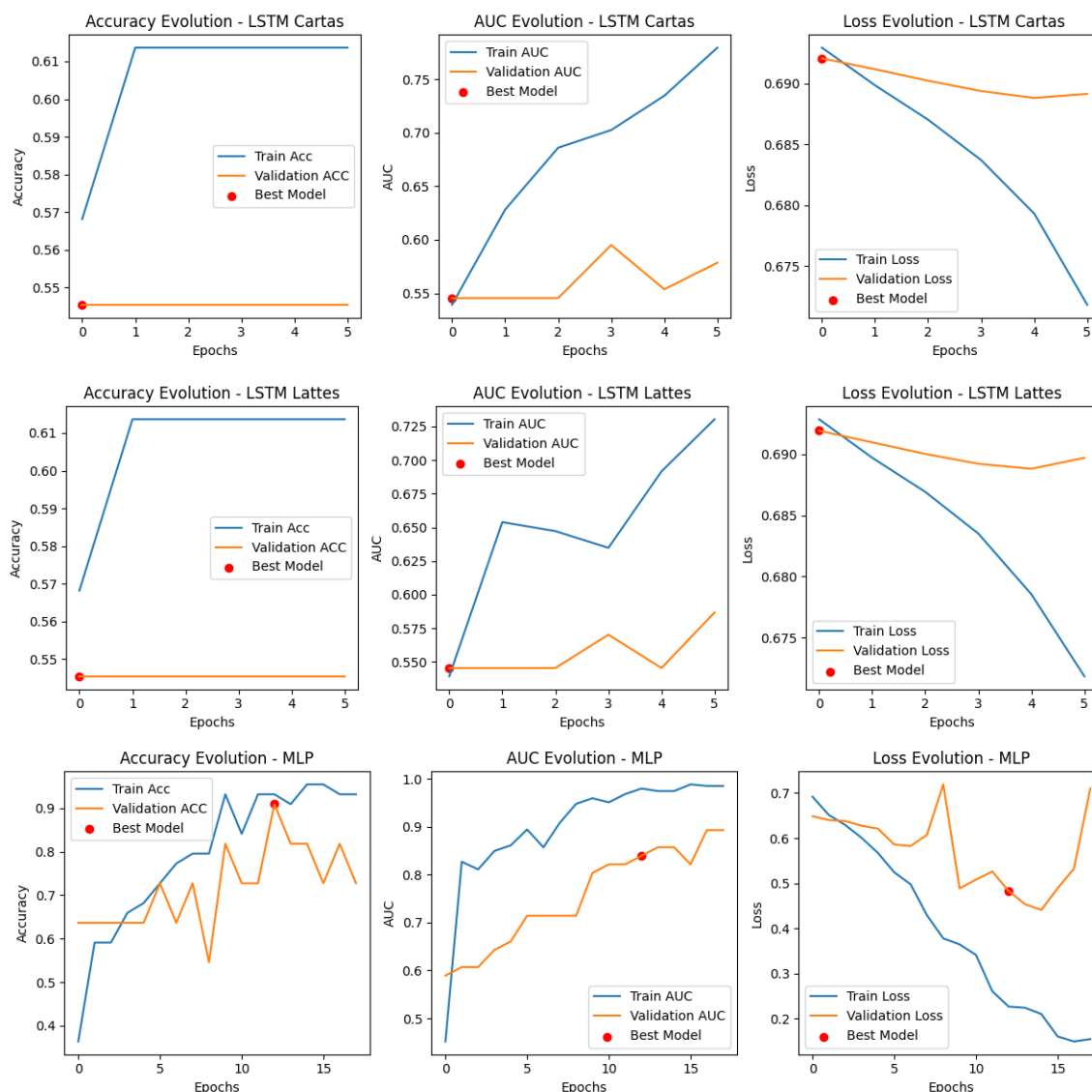
Estes resultados evidenciam não só a viabilidade das técnicas de aprendizado de máquina em contextos educacionais, mas também a necessidade de seleção criteriosa dos modelos de acordo com a natureza dos dados disponíveis. A pesquisa reforça a importância de continuar a otimização de modelos de aprendizado de máquina, explorando sua aplicação prática na seleção acadêmica e no monitoramento do desempenho dos estudantes, com implicações significativas para a eficiência dos processos seletivos e para a estratégia de acompanhamento acadêmico.

5 CONSIDERAÇÕES FINAIS

O estudo em questão investigou a aplicabilidade das cartas de intenção e informações do Currículo *Lattes* como preditores do desempenho acadêmico, utilizando para tal, modelos avançados de aprendizado de máquina, especificamente redes neurais LSTM e MLP. Os resultados obtidos, apesar de variados, oferecem *insights* valiosos para o aprimoramento da seleção de candidatos a programas de especialização, assim como para o desenvolvimento de estratégias pedagógicas baseadas em dados.

As redes LSTM, apresentaram um desempenho moderado, sugerindo uma boa capacidade de identificar todos os candidatos de sucesso, mas com limitações na discriminação entre os diferentes níveis de desempenho dos candidatos. A variabilidade na área sob a curva ROC entre as cartas de intenção e os Currículos *Lattes* aponta para uma capacidade discriminativa inconsistente, indicando a necessidade de otimizar a seleção de dados e a configuração do modelo. Comparativamente, as LSTM foram menos eficazes que a rede MLP, destacando a importância

Figura 5 – Métricas apresentadas pelas redes neurais



Fonte: Elaborado pelo autor

da escolha do modelo adequado para a tarefa de previsão em questão.

A rede MLP mostrou-se promissora, superando as redes LSTM em termos de acurácia e precisão. Isso não só confirma a utilidade do aprendizado de máquina em contextos educacionais, mas também destaca a importância de continuar a explorar e otimizar diferentes arquiteturas de rede para tarefas de previsão.

A relevância destas descobertas é dupla: por um lado, valida a integração de métodos de aprendizado de máquina no ensino superior, especialmente na seleção de candidatos; por outro, ressalta a necessidade de coleta e análise cuidadosa de dados educacionais. As implicações práticas são consideráveis, potencialmente contribuindo para a redução das taxas de evasão e melhoria do acompanhamento do desenvolvimento estudantil.

Em suma, este estudo contribui significativamente para o campo educacional, fornecendo evidências concretas da eficácia do uso de técnicas de aprendizado de máquina na seleção de candidatos. Incentiva-se a continuidade da pesquisa nesta área, buscando sempre a otimiza-

ção dos modelos e a exploração de novas variáveis e conjuntos de dados, para robustecer ainda mais as estratégias de seleção e acompanhamento acadêmico.

REFERÊNCIAS

- ALMEIRA, L. S.; CASANOVA, J. R. Desenvolvimento psicossocial e sucesso academico no ensino superior. in f. veiga (coord.),. **Psicologia da Educação: Temas de Aprofundamento Científico para a Educação XXI**, Climepsi Editores, Lisboa, p. 101–128, 2019.
- ALOM, M. Z.; TAHA, T. M.; YAKOPCIC, C.; WESTBERG, S.; SIDIKE, P.; NASRIN, M. S.; HASAN, M.; ESSEN, B. C. V.; AWWAL, A. A. S.; ASAR and Vijayan K. A state-of-the-art survey on deep learning theory and architectures. **electronics**, v. 8, n. 3, p. 292, 2019.
- BRASIL. Lei nº 13.709, de 14 de agosto de 2018. **Diário Oficial [da] República Federativa do Brasil**, Brasília, DF, 2018. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>.
- BRITO, A. G. C. de; QUONIAM, L.; MENA-CHALCO, J. P. Exploração da plataforma lattes por assunto: proposta de metodologia. **Transinformação**, v. 28, p. 77–86, 2016.
- CASIRAGHI, B.; BORUCHOVITCH, E.; ALMEIDA, L. d. S. Variáveis psicológicas e seu impacto no rendimento acadêmico no ensino superior. **Revista Brasileira de Educação**, n. 27, 2022.
- CHANG, C.; MASTERSON, M. Using word order in political text classification with long short-term memory models. **Political Analysis**, Cambridge University Press, v. 28, n. 3, p. 395–411, 2020.
- COLPO, M. P.; PRIMO, T. T.; PERNAS, A. M.; CECHINEL, C. Mineração de dados educacionais na previsão de evasão: uma rsl sob a perspectiva do congresso brasileiro de informática na educação. **Anais do XXXI Simpósio Brasileiro de Informática na Educação**, Porto Alegre, p. 1102–1111, 2020.
- DANTAS, S. de O. Cartas de intenção em processos seletivos acadêmicos: efeitos de sentido sobre o pet letras ufu e a imagem de aluno-candidato. **Dissertação (Mestrado em Estudos Linguísticos)**, 2022.
- EMMERT-STREIB, F.; YANG, Z.; FENG, H.; TRIPATHI, S.; DEHMER, M. An introductory review of deep learning for prediction models with big data. **Applied sciences**, v. 3, 2020.
- GARCIA, L. M. L. S.; LARA, D. F.; ANTUNES, F. Análise da retenção no ensino superior: um estudo de caso em um curso de sistemas de informação. **Revista da Faculdade de Educação**, v. 34, n. 2, p. 15–38, 2020.
- GE, W.; LIN, X.; YU, Y. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. p. 3034–3043, 2019.
- JANG, B.; KIM, M.; HARERIMANA, G.; KANG, S.-u.; KIM, J. W. Bi-lstm model to increase accuracy in text classification: Combining word2vec cnn and attention mechanism. **Applied Sciences**, v. 10, n. 17, 2020. ISSN 2076-3417. Disponível em: <<https://www.mdpi.com/2076-3417/10/17/5841>>.

LEDIG, C.; THEIS, L.; HUSZAR, F.; CABALLERO, J. Photo-realistic single image super-resolution using a generative adversarial network. **Proceedings of the IEEE conference on computer vision and pattern recognition**, 2017.

LIU, G.; GUO, J. Bidirectional lstm with attention mechanism and convolutional layer for text classification. **Neurocomputing**, v. 337, p. 325–338, 2019. ISSN 0925-2312. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0925231219301067>>.

LOPES, R. C. Currículo lattes: um estudo sobre algumas representações sociais. **PragMATIZES-Revista Latino-Americana de Estudos em Cultura**, v. 12, n. 22, p. 663–694, 2022.

MARIANO, A. M.; SANTOS, M. R. Revisão da literatura: Apresentação de uma abordagem integradora. **AEDem International Conference**, p. 427–442, 2017.

MUNIR, K.; ELAHI, H.; AYUB, A.; FREZZA, F.; RIZZI, A. Cancer diagnosis using deep learning: a bibliographic review. **Cancers**, v. 11, n. 9, p. 1235, 2019.

NOBRE, L. N.; FREITAS, R. R. de. A evolução da pós-graduação no brasil: histórico, políticas e avaliação. **Brazilian Journal of Production Engineering**, v. 3, n. 2, p. 18–30, 2017.

ORUH, J.; VIRIRI, S.; ADEGUN, A. Long short-term memory recurrent neural network for automatic speech recognition. **IEEE Access**, v. 10, p. 30069–30079, 2022.

RASTROLLO-GUERRERO, J. L.; GOMEZ-PULIDO, J. A.; DURAN-DOMINGUEZ, A. Analyzing and predicting students' performance by means of machine learning: A review. **Applied sciences**, v. 10, n. 3, p. 1042, 2020.

ROMERO, C.; VENTURA, S. Educational data mining and learning analytics: An updated survey. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 10, n. 3, 2020.

SAGHEER, A.; MOSTAFA, K. Unsupervised pre-training of a deep lstm-based stacked autoencoder for multivariate time series forecasting problems. **Scientific reports**, v. 9, n. 1, p. 1–16, 2019.

SCHMIDHUBER, J. Deep learning in neural networks: An overview. **Neural networks**, v. 61, p. 85–117, 2015.

SILVA, A.; VIEIRA, N. Mapeamento longitudinal da retenção e evasão em cursos de licenciatura como contribuição à governança educacional: um estudo de caso no instituto federal de minas gerais. **Ensino e Pesquisa, União da Vitória**, v. 20, n. 2, p. 54–68, 2022.

SOUZA, F. G. D. C. Previsão de evasão e retenção escolar no ensino médio profissional: uma abordagem baseada em redes neurais. artificiais. dissertação (mestrado) - programa de pós-graduação em educação profissional e tecnológica. Salgueiro, 2021.

SOUZA, V. F. de; CAZELLA, S. C. Mineração de dados educacionais com algoritmos de regressão: um estudo sobre a predição do desempenho. **Revista Educar Mais**, v. 6, p. 183–198, 2022.

SU, C.; HUANG, H.; SHI, S.; JIAN, P.; SHI, X. Neural machine translation with gumbel tree-lstm based encoder. **Journal of Visual Communication and Image Representation**, v. 71, p. 102811, 2020. ISSN 1047-3203. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1047320320300614>>.

SUTSKEVER, I.; VINYALS, O.; LE, Q. V. Sequence to sequence learning with neural networks. **Advances in neural information processing systems**, v. 27, 2014.

TOMASEVIC, N.; GVOZDENOVIC, N.; VRANES, S. An overview and comparison of supervised data mining techniques for student exam performance prediction. **Computers and Education**, v. 143, p. 103676, 2020.

WAHEED, H.; HASSAN, S.-U.; ALJOHANI, N. R.; HARDMAN, J.; ALELYANI, S.; NAWAZ, R. Predicting academic performance of students from vle big data using deep learning models. **Computers in Human Behavior**, v. 104, p. 106189, 2020.

XU, G.; MENG, Y.; QIU, X.; YU, Z.; WU, X. Sentiment analysis of comment texts based on bilstm. **IEEE Access**, v. 7, p. 51522–51532, 2019.

YONG, Y.; XIAOSHENG, S.; CHANGHUA, H.; JIANXUN, Z. A review of recurrent neural networks: Lstm cells and network architectures. **Neural Computation**, v. 31, n. 7, p. 1235–1270, 2019.

ZHANG, J.; LIU, F.; XU, W.; YU, H. Feature fusion text classification model combining cnn and bigru with multi-attention mechanism. **Future Internet**, v. 11, n. 11, 2019. ISSN 1999-5903. Disponível em: <<https://www.mdpi.com/1999-5903/11/11/237>>.