Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: 07/Sep/2023; 1st round notif.: 04/Dec/2023;
Camera ready: 22/Jan/2024; Edition review: 11/Apr/2024; Available online: 13/Apr/2024; Published: 13/Apr/2024;

Transformers para previsão de desempenho acadêmico no ensino Fundamental e Médio

Title: Student Performance Transformer in Primary and Secondary education

Título: Transformadores para predecir el rendimiento académico en educación Primaria y Secundaria

Lorran Santos Rodrigues
Instituto Militar de Engenharia
ORCID:0000-0003-1031-1323
lorranrodr@ime.eb.br

Marcos Santos
Instituto Militar de Engenharia
ORCID:0000-0003-1533-5535
marcosdossantos@ime.eb.br

Carlos Francisco Simoes Gomes
Universidade Federal Fluminense
ORCID:0000-0002-6865-0275
cfsg1@bol.com.br

Ricardo Choren
Instituto Militar de Engenharia
ORCID:0000-0003-4081-2647
choren@ime.eb.br

Ronaldo Goldschmidt
Instituto Militar de Engenharia
ORCID:0000-0003-1688-0586
ronaldo.rgold@ime.eb.br

Saulo Barbará
Universidade Federal Rural do Rio de Janeiro
ORCID:0000-0002-9424-5425
saulobarbara@gmail.com

Resumo

A previsão de desempenho acadêmico apresenta um potencial grande no trabalho pró-ativo das escolas na identificação de alunos em risco de reprovação. de duas redes distintas, permitindo a comparação entre diferentes anos escolares, anos letivos e redes de ensino. Contrastaram-se os desempenhos de modelos baseados na arquitetura Transformers com modelos mais estabelecidos, como o XGBoost e um modelo de rede neural mais simples. Os resultados mostraram que os Transformers tiveram um desempenho interessante na tarefa de previsão de desempenho acadêmico, especialmente com um número maior de avaliações. No entanto, o XGBoost conseguiu alcançar um alto desempenho mais cedo no período letivo. Uma vantagem dos Transformers é sua flexibilidade no treinamento, permitindo lidar com conjuntos de dados semi-estruturados sem a necessidade de pré-processamento. Em última análise, esta pesquisa contribui para o desenvolvimento de métodos que podem identificar precocemente alunos em risco de reprovação, oferecendo a oportunidade de intervenção e apoio adequados. Isso pode ter um impacto positivo na formação dos alunos e na sociedade como um todo, mitigando prejuízos e promovendo a educação de qualidade.
Palavras-chave: *Desempenho acadêmico; Transformers; EDM.*

Abstract

Academic performance prediction holds significant potential for proactive school efforts in identifying students at risk of failing. This study was motivated by the gap in academic performance prediction studies focused on middle and elementary education that explore deep learning models. This led to the collection of a dataset of basic education students in mathematics and Portuguese from two distinct networks, allowing for comparisons across different school years, academic years, and educational networks. The performances of models based on Transformer architectures were contrasted with more established models such as XGBoost and a simpler neural network model. The results showed that Transformers had an interesting performance in the academic performance prediction task, especially with a larger number of assessments. However, XGBoost managed to achieve high performance earlier in the school term. An advantage of Transformers is their flexibility in training, allowing them to handle semi-structured datasets without the need for preprocessing. Ultimately, this research contributes to the development of methods that can identify students at risk of failing early, offering the opportunity for timely intervention and support. This can have a positive impact on students' education and society as a whole, mitigating losses and promoting quality education.

Cite as: Rodrigues, LS, Santos, M, Gomes, CFS, Choren, R, Goldschmidt, R, Barbará, Saulo (2024). Transformers para previsão de desempenho acadêmico no ensino Fundamental e Médio. Revista Brasileira de Informática na Educação, 32, 213-241. https://doi.org/10.5753/rbie.2024.3661.

Keywords: *Academic performance; Transformers; EDM.*

Resumen

La predicción del rendimiento académico tiene un gran potencial en los esfuerzos proactivos de las escuelas para identificar a los estudiantes en riesgo de fracaso. Este estudio fue motivado por la brecha en los estudios de predicción del rendimiento académico centrados en la educación media y primaria que exploran modelos de aprendizaje profundo. Esto llevó a la recopilación de un conjunto de datos de estudiantes de educación básica en Matemáticas y Portugués de dos redes educativas distintas, lo que permitió comparaciones entre diferentes años escolares, años académicos y redes educativas. Se contrastaron los desempeños de modelos basados en arquitecturas Transformer con modelos más establecidos como XGBoost y un modelo de red neuronal más simple. Los resultados mostraron que los Transformers tuvieron un desempeño interesante en la tarea de predicción del rendimiento académico, especialmente con un mayor número de evaluaciones. Sin embargo, XGBoost logró alcanzar un alto rendimiento más temprano en el período escolar. Una ventaja de los Transformers es su flexibilidad en el entrenamiento, lo que les permite manejar conjuntos de datos semi-estructurados sin necesidad de preprocesamiento. En última instancia, esta investigación contribuye al desarrollo de métodos que pueden identificar tempranamente a los estudiantes en riesgo de fracaso, ofreciendo la oportunidad de intervención y apoyo oportunos. Esto puede tener un impacto positivo en la educación de los estudiantes y en la sociedad en su conjunto, mitigando pérdidas y promoviendo una educación de calidad.

Palabras clave: *Desempeño académico; Transformers; EDM.*

1 Introdução

A mineração de dados educacionais (*educational data mining*, EDM) se preocupa, principalmente, em desenvolver, pesquisar e aplicar técnicas de aprendizado de máquina, mineração de dados e outros métodos estatísticos para detectar padrões em grandes coleções de dados educacionais (Romero et al., 2010). Neste campo, de acordo com Hernández-Blanco et al. (2019), o desafio que mais tem chamado atenção é a previsão de desempenho acadêmico do aluno.

Essa tarefa, em suas mais variadas formas, tomou considerável atenção principalmente no que diz respeito ao contexto do ensino superior (Zhang & Li, 2018). Prever qual estudante pode apresentar um desempenho baixo, o mais rápido possível, é estratégico para instituições de ensino que passam a agir de forma preventiva para sanar as possíveis deficiências e garantir a melhor experiência do aluno (Namoun & Alshanqiti, 2021). Sob outro aspecto, a previsão de desempenho acadêmico pode também corroborar com a descoberta de alunos com alto potencial (Tatar & Dütegör, 2020).

A literatura recente apresentou bons resultados ao modelar o problema de previsão de desempenho do aluno como um problema de previsão sequencial utilizando aprendizado profundo (Hussain et al., 2019; Kim et al., 2018). No entanto, esse tipo de modelagem ainda não foi explorada em um cenário com dados não provenientes de sistemas de aprendizado – MOOC ou outras plataformas de ensino digital –, (Rodrigues et al., 2022). Além disso, com base na revisão da literatura apresentada por (Rodrigues et al., 2022), existe uma lacuna de trabalhos que utilizam aprendizado profundo para previsão de desempenho acadêmico no ensino fundamental e médio, mais especificamente modelos que lançam mão da arquitetura proposta por Vaswani et al. (2017). Soma-se isso a uma baixa presença de trabalhos que se utilizam de conjuntos de dados grandes para realizar a validação de seus modelos (Namoun & Alshanqiti, 2021). Estes fatores, portanto,

serviram como motivadores para o presente trabalho.

Logo, como objetivo principal da pesquisa, pretende-se estudar a utilização de redes neurais de aprendizado profundo (DNN), utilizando-se a arquitetura *Transformer*, para previsão do desempenho acadêmico, representado como a aprovação ou reprovação do aluno no período letivo, no contexto do ensino médio e fundamental. Além disso, se objetiva o treinamento do modelo em dados semi-estruturados referente a sequência, de tamanho variável, de avaliações realizadas pelos alunos. Essa modelagem tende a ser particularmente interessante, uma vez que possibilita a ingestão de dados para o modelo sem necessidade de pré-processamento, ao contrário de outros modelos clássicos de aprendizado de máquina. Isto porque escolas diferentes podem dispor de esquemas de avaliações diferentes, como provas bimestrais, trimestrais, semestrais ou com intervalos não definidos, e nomenclaturas de identificação distintas, que no caso de modelos tradicionais pode representar um esforço extra e recorrente na adaptação do modelo para cada escola. Atrelar um modelo para previsão de desempenho acadêmico a um esquema de avaliações pré-definido, torna o modelo gerado menos flexível, portanto, potencialmente de adoção com mais atrito pelas instituições de ensino interessadas.

Após o modelo treinado, se espera fazer a previsão do desempenho dos alunos tendo-se como entrada os dados, semi-estruturados, de notas em avaliações periódicas. O modelo proposto será avaliado em um conjunto de dados proveniente de múltiplas escolas, pertencentes a duas redes de ensino, de forma que seja possível se fazer a comparação do modelo aplicado a diferentes populações.

As questões de pesquisa que este trabalho se propõem a responder são:

- Questão 1: Existe benefício em se utilizar aprendizado profundo, baseado na arquitetura *Transformer*, para previsão de desempenho acadêmico em dados semi-estruturados de alunos do ensino fundamental e médio?
- Questão 2: Caso haja benefício, a partir de qual ponto ele se mostra significativo, considerando o período letivo?

A hipótese levantada foi a seguinte:

- Hipótese 1: Os modelos de aprendizado profundo, baseado na arquitetura *Transformer*, treinados em dados semi-estruturados de alunos do ensino fundamental e médio para previsão de desempenho acadêmico possuem uma eficiência na tarefa de classificação superior aos modelos clássicos de aprendizado de máquina.

2 Referencial Teórico

Nesta seção serão exploradas as bases conceituais e teóricas que sustentam a investigação em questão. Este segmento oferecerá uma análise abrangente das teorias e estudos relevantes que fornecem o alicerce para a compreensão do problema de pesquisa abordado. Ao explorar as contribuições acadêmicas e as perspectivas existentes, pretende-se situar o estudo no contexto mais amplo da literatura científica, identificando lacunas de conhecimento e estabelecendo conexões cruciais entre as ideias preexistentes e a abordagem proposta.

2.1 Descoberta de conhecimento em Bases de Dados

O processo de Descoberta de conhecimento em Bases de dados (KDD) é constituido de três etapas: pré-processamento, mineração de dados e pós-processamento, sendo esta última crucial (Goldschmidt et al., 2015). A aplicação de KDD no contexto de dados educacionais resultou em um novo campo, a mineração de dados educacionais (EDM) (Hernández-Blanco et al., 2019). Uma das ferramentas utilizadas no processo de KDD são modelos de aprendizado de máquina (Goldschmidt et al., 2015)

Bonaccorso (2017) define primeiro aprendizado, para em seguida definir o que seria aprendizado de máquina. O autor define aprendizado como a habilidade de se modificar de acordo com estímulos externos e se recordar das experiências passadas. Nesse sentido, aprendizado de máquina seria uma abordagem de engenharia que foca em técnicas que aumentam ou melhoram a propensão de se modificar de forma adaptativa. Definições mais formais foram dadas por dadas por Mitchell et al. (2007) e Goodfellow et al. (2016). Geralmente, A tarefa de aprendizado de máquina é descrita a partir da forma em que o sistema responsável por aprender a tarefa deve processar um exemplo, ou amostra (Goodfellow et al., 2016).

2.2 Redes Neurais

Redes neurais artificiais (RNA) podem ser entendidas como sistemas de computação inspirados, abstratamente, pelas redes neurais biológicas, que ocorrem naturalmente no cérebro animal (Livingstone, 2008). Redes neurais artificiais são compostas por unidades de processamento, chamadas neurônios artificiais. O primeiro tipo de neurônio usado em uma RNA foi proposto por Rosenblatt (1958). Chamado de *perceptron*, ele possui uma arquitetura simples e se tornou obsoleto para resolução de problemas não lineares devido a sua inflexibilidade e falta de estabilidade de sua função de ativação ($g(.)$). No caso simples de uma saída binária, a função de ativação é responsável por dizer se neurônio será ativado ou não (Sharma et al., 2017).

As redes *Perceptron* Multicamadas (*Multi-layer Perceptron* ou MLP) são caracterizadas pela presença de múltiplas camadas na unidade básica de processamento do tipo *Perceptron* (Guyon, 1991). A adição de camadas impõe um aumento na capacidade de processamento não linear e generalização da rede como um todo. As redes MLP também são conhecidas como redes *feedforward*. O modelo gerado é associado com um grafo acíclico dirigido que descreve como as funções que compõem o modelo são arranjadas. Por exemplo, tendo-se três funções $f^{(1)}$, $f^{(2)}$ e $f^{(3)}$ encadeadas, para formar $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$. Goodfellow et al. (2016) afirmam que essa estrutura encadeada é a estrutura mais comum em redes neurais. Nesse exemplo $f^{(1)}$ representa a primeira camada, $f^{(2)}$ a segunda camada, assim sucessivamente. O tamanho da cadeia define a profundidade do modelo. A camada final da rede também é conhecida como camada de saída.

De acordo com Goodfellow et al. (2016) as redes neurais, *feedforward* são de grande importância no campo de aprendizado de máquina. Elas formam a base de muitas aplicações comerciais modernas, como, por exemplo, as redes convolucionais utilizada para detecção de objetos em imagens, as quais são um tipo especializado de rede *feedforward*.

2.3 Modelos Sequenciais

Como grande parte do progresso dos modelos sequências se deu a partir do campo de processamento de linguagem natural (PLN), se faz necessário apresentar o desenvolvimento de tais modelos a partir da perspectiva desse campo específico de pesquisa. Um problema comum em PLN seria tentar inferir a partir de um corpo de texto, uma classe. Ex.: A partir do conteúdo de um e-mail identificar se ele é ou não um *spam*. Sucessivas evoluções se deram nas arquiteturas de redes neurais, para possibilitar o trabalho com a entrada de dados sequencial, como é comum em problemas textuais, essas evoluções passaram pelo desenvolvimento de redes neurais recorrentes (RNN) a redes neurais de memória de longo e curto prazo (LSTM). Para uma revisão a respeito dessas arquiteturas consulte Goodfellow et al. (2016) e Xiao e Zhou (2020).

O destaque mais recente é dado ao desenvolvimento do mecanismo de atenção proposto por Bahdanau et al. (2014). O mecanismo em questão foi responsável por permitir o desenvolvimento da arquitetura *Transformer* proposta por sua vez por Vaswani et al. (2017) para solucionar um problema de tradução automatizada.

A maneira clássica de se trabalhar com problema de tradução automatizada (ex.: tradução de um texto em inglês para o francês) em redes neurais é a partir de modelos *seq2seq*, que se trata de uma arquitetura que permite tomar uma sequência de entrada e fornecer uma sequência como saída (Goodfellow et al., 2016). Esses modelos são compostos por dois blocos: um codificador e um decodificador. O codificador é encarregado de processar cada item na sequência, compilando a informação em um vetor (chamado, contexto). Após processar a entrada, o codificador envia esse contexto para o decodificador que produz a saída, item por item. Ambos o codificador e o decodificador, em um primeiro momento, são redes neurais recorrentes, ou seja, processam sequencialmente o conjunto de entrada (Goodfellow et al., 2016).

Foi sugerido por Bahdanau et al. (2014) um mecanismo, que permitia ao decodificador revisitar a sequência de entrada a cada etapa do processamento, ou seja, a cada nova entrada sendo processada se torna possível observa-la no contexto onde está inserida. Ademais, ao invés de receber a mesma representação da entrada, foi proposto que o decodificador pudesse, de maneira seletiva, focar em partes particulares da sequência de entrada, em diferentes etapas do processo de decodificação. Nesse sentido, o mecanismo de atenção de Bahdanau et al. (2014) proporcionou uma forma do decodificador observar partes diferentes da entrada a cada etapa de decodificação.

Em outras palavras, o codificador poderia produzir uma representação de tamanho igual à sequência original. Então, na etapa de decodificação o decodificador poderia (via um mecanismo de controle) receber um vetor de contexto que consiste na soma ponderada da representação na entrada de cada etapa. Intuitivamente, os pesos determinariam o quão cada etapa deve focar em cada valor de entrada, e a chave seria fazer esse processo de atribuição de pesos diferenciável, para que possam ser aprendidos em conjunto com outros parâmetros da rede (Bahdanau et al., 2014).

Em suma, os mecanismos de atenção emergiram com relevância, superando sua utilidade de aprimoramento de modelos *seq2seq*. Vaswani et al. (2017) propuseram a arquitetura *Transformer* dispensando as RNN, por completo, e dependendo apenas de mecanismos de atenção para capturar os relacionamentos entre entrada e saída. A arquitetura apresentou excelente desempenho e em 2018 os modelos baseados em *Transformer* passaram a figurar em grande parte no estado da arte em processamento de linguagem natural, principalmente por permitirem um nível elevado

de paralelismo em GPUs, justamente por não dependerem de RNN, ou seja dispensando o trabalho de processamento sequencial. Uma revisão a respeito do mecanismo de atenção, e suas aplicabilidades para além da PLN foi proposta em Niu et al. (2021).

2.4 Previsão de Desempenho Acadêmico

De acordo com Márquez-Vera et al. (2013), a previsão de desempenho acadêmico tem se tornado de grande importância para as instituições de ensino. Guo et al. (2016) exemplificam que ao fornecer o aviso devidamente antecipado de um estudante em risco, possibilita que as coordenações pedagógicas se mobilizem proativamente para auxiliar o aluno a superar suas dificuldades em determinada disciplina.

A problemática da previsão de desempenho acadêmico pode se dar em múltiplos objetivos (Hellas et al., 2018; Liz-Domínguez et al., 2019; Namoun & Alshanqiti, 2021). Os mais recorrentes na literatura possuem como objetivo prever a nota de um aluno ao final de um período de avaliação (Fernandes et al., 2019), ou prever em qual faixa de nota o aluno estará ao fim do período (A, B, C, D e F ou Aprovado / Reprovado). No entanto, não é incomum encontrar trabalhos prevendo o desempenho acadêmico como uma nota de uma prova específica, ou a probabilidade de retenção/evasão de um aluno (Rovira et al., 2017). Esta falta de definição clara é problemática, e já foi explicitada por Hellas et al. (2018). Por esta razão é importante ressaltar que o presente trabalho se atêm à previsão do desempenho acadêmico como a classificação se um aluno será aprovado ou reprovado ao final do período letivo.

Os conjuntos de dados explorados pela EDM são variados. No entanto, existe um crescente interesse em dados oriundos de plataformas de educação virtual. Principalmente de cursos *online* abertos e massivos (MOOC), como reportado por Hussain et al. (2019) e também apresentado por Gottardo et al. (2014). No contexto dessas inovações, tornou-se possível obter dados de interações de alunos com a instituição de ensino de uma forma mais intensa (ex.: visualização de arquivos, tempo de leitura de material, quantidade de visitas a determinado recurso, etc.). No entanto, os trabalhos levantados que exploram a previsão de desempenho acadêmico no ensino médio e fundamental, utilizam conjuntos mais "clássicos" de dados (Rodrigues et al., 2022). Ou seja, dados pertencentes a dimensão escolar do aluno, como nota de provas e presença em aulas; dados pertencentes a dimensão familiar, como profissão dos pais, estado civil, número de coabitantes, etc.; dados pertencentes à dimensão pessoal do aluno, como sexo, etnia e idade. Embora alguns estudos reportem a utilidade da inclusão de dados socio-demográficos (Barros et al., 2019), o desempenho acadêmico geralmente é bem descrito a partir das dimensões escolares (Almayan & Al Mayyan, 2016; Imran et al., 2019; Roy & Garg, 2017; Souza & Santos, 2021).

3 Trabalhos Relacionados

A Tabela 1 foi retirada da revisão sistemática proposta por Rodrigues et al. (2022). Nela tem-se um conjunto de trabalhos recentes que utilizam aprendizado de máquina para previsão de desempenho acadêmico de alunos do ensino fundamental e médio. No trabalho de revisão foi identificado que nenhum dos trabalhos utilizou uma arquitetura de rede baseada em *transformers* para previsão de desempenho acadêmico, além disso, nenhum deles utiliza-se de um conjunto de dados semi-

estruturados para o treinamento do modelo.

Trabalho	**T?**	**DSE?**
Márquez-Vera et al., 2013	0	0
Guo et al., 2016	0	0
Almayan e Al Mayyan, 2016	0	0
Blasi, 2017	0	0
Amra e Maghari, 2017	0	0
Fernandes et al., 2019	0	0
Roy e Garg, 2017	0	0
Athani et al., 2017	0	0
Yang e Li, 2018	0	0
Zaffar et al., 2018	0	0
Lu e Yuan, 2018	0	0
Qazdar et al., 2019	0	0
Imran et al., 2019	0	0
Lee e Chung, 2019	0	0
Barros et al., 2019	0	0
Turabieh, 2019	0	0
Gil et al., 2020	0	0
Orooji e Chen, 2019	0	0
García-González e Skrita, 2019	0	0
Cornell-Farrow e Garrard, 2020	0	0
Razaque e Alajlan, 2020	0	0
Sokkhey e Okazaki, 2020	0	0
H. Alamri et al., 2020	0	0
Kim et al., 2018	0	0
Q. Chen et al., 2019	0	0
Hussain e Khan, 2021	0	0

Tabela 1: Artigos avaliados e suas características. Utilizam a arquitetura Transformer (T) e Utilizam Dados Semi-Estruturados (DSE) Fonte: Adaptado de (Rodrigues et al., 2022).

Ainda assim, ressalta-se que outros dois trabalhos se relacionam com o trabalho presente. Eles são os trabalhos propostos por Kim et al. (2018) e W. Chen et al. (2019).

No trabalho de (Kim et al., 2018) foi proposto um modelo de aprendizado profundo que consumia uma sequência de eventos dos alunos, de uma plataforma de aprendizado *online*, e efetuava inferências a respeito da aprovação dos mesmos no curso. O conceito do consumo da sequência de eventos dos alunos é particularmente interessante também para conjuntos de dados de ensino médio e fundamental por permitir a utilização de um modelo de aprendizado profundo a partir de um conjunto de dados de um cenário genérico. Uma vez que escolas possuem autonomia para determinarem a quantidade de avaliações ao longo do ano, depender de mecanismos de pré-processamento de dados para aplicação de modelos de inferência pode representar um risco a sistemas que utilizam tais modelos.

Soma-se o crescente interesse em modelos de aprendizado profundo que utilizam o meca-

nismo de atenção proposto por Vaswani et al. (2017). Entendeu-se que uma adaptação para o caso apresentado seria proveitoso.

Além do trabalho de Kim et al. (2018), em Q. Chen et al. (2019) foi proposta utilização da arquitetura *Transformer* para sistemas de recomendação, por dados de entrada que partiam de uma sequência. Os autores afirmam que essa arquitetura é interessante, pois para o problema de recomendação de itens em um catálogo de *e-commerce* o aspecto da sequência de itens sendo escolhidos é também importante, além dos próprios itens em si. Este aspecto da sequência também é relevante para o problema de previsão de desempenho acadêmico, pois a partir dela se espera também capturar as características de adaptação de um aluno as matérias e a evolução do mesmo.

Tendo-se em vista o sucesso dos modelos *Transformer* para tarefas de tradução em processamento de linguagem natural – (Devlin et al., 2018; Vaswani et al., 2017) – foi proposta a aplicação de mecanismo de atenção a fim de se criar uma melhor representação de cada avaliação em uma sequência, ao longo de um período letivo, para determinar o desempenho acadêmico de alunos. Esta informação sequencial foi considerada na etapa de *embedding* e em seguida alimenta o bloco *transformer* buscando inferir se o aluno será ou não aprovado ao fim do período letivo.

O diferencial da arquitetura *Transformer no campo de processamento de linguagem natural reside em sua habilidade de capturar a dependência entre as palavras em uma frase, mediante o mecanismo de atenção. De acordo com as observações de Q. Chen et al. (2019), de maneira intuitiva, a dependência entre elementos em uma sequência, não necessariamente palavras, também poderia ser discernida por meio de um modelo Transformer. Portanto, a extensão sugerida neste estudo foca agora na aplicação desse conceito no contexto da previsão de desempenho acadêmico.*

Assim, sugeriu-se treinar um modelo que se baseia na arquitetura *Transformer* (Vaswani et al., 2017), mais especificamente uma adaptação do que foi trabalhado em Q. Chen et al. (2019) para se trabalhar com uma sequência genérica de avaliações e entender sua aplicabilidade na tarefa de previsão do desempenho acadêmico. Além disso, lançou-se mão de outros modelos para se contrastar os resultados obtidos.

4 Conjunto de Dados

De modo a realizar o experimento proposto, foi obtido um conjunto de dados de duas redes de ensino, com cerca de 16 escolas cada, com atuação em um estado brasileiro. Ao todo, coletou-se o histórico de avaliações em duas disciplinas, sendo elas, Português e Matemática ao longo de dois anos letivos. A base contem avaliações de 5792 alunos distintos do primeiro ano do ensino médio e do 9ž ano do ensino fundamental.

Não foi coletada nenhuma informação que identificasse os alunos ou as redes, ou escolas sendo avaliadas. Na verdade, os dados foram obtidos já de forma anonimizada, de forma que ainda fosse possível distinguir as respectivas redes e turmas, porém não identificá-las nominalmente.

A base de dados contêm 234996 registros, sendo cada linha do conjunto de dados uma avaliação, de um aluno, de uma escola, em uma disciplina, em um ano escolar (ou série escolar) específico, durante um período letivo. As avaliações são numeradas com sua devida ordem de

aplicação no período letivo. Os dados referentes aos anos de 2018 e 2019 foram adquiridos, considerando os impactos significativos gerados pela pandemia de COVID-19 nas instituições de ensino no Brasil a partir de 2020. A escolha desses anos visa mitigar os efeitos abruptos que os eventos pandêmicos podem exercer sobre a avaliação do desempenho acadêmico em condições normais de ensino presencial.

Logo, tem-se um conjunto de dados: diverso, com duas populações, uma pertencente a cada rede de ensino, em anos distintos. Além disso, expressivo, em contraste com trabalhos avaliados na revisão de Rodrigues et al. (2022) que utilizam o conjunto de dados disponibilizado por Cortez e Silva (2008), que, se trata de um conjunto de dados, antigo, com poucos alunos (649), de uma única escola, porém com um grande número de atributos (33).

Foi utilizada a terminologia "série" para se referir aos anos escolares estudados, no caso o 9ž do ensino fundamental (série:3) e o 1ž ano (série:21) do ensino médio, para evitar a ambiguidade com a terminologia de ano letivo.

Ambas as escolas adotam um sistema de avaliações com peso. Ou seja, avaliações diferentes podem possuir pesos diferentes, exemplo: uma prova bimestral pode representar um peso maior para avaliação final do aluno do que uma avaliação optativa. As notas obtidas no conjunto de dados já foram normalizadas pelo peso. Logo, tem-se para cada avaliação o quanto cada aluno atingiu da pontuação do total possível naquela avaliação.

Embora pertencentes ao mesmo grupo de ensino, as escolas possuem autonomia para determinar suas respectivas formas de avaliação. Assim, as escolas possuem a maioria das avaliações em comum, no que tange ao objetivo de avaliação (conteúdo programático a ser avaliado), porém algumas específicas a sua própria metodologia de ensino.

Além da nota em cada avaliação, existe também um atributo, replicado para cada registro no conjunto de dados, que informa se o aluno foi aprovado ou não naquela disciplina específica ao fim do período letivo.

É importante apontar que no caso da disciplina Matemática, para o primeiro ano do ensino médio de ambas as redes, cada avaliação é referente a um subconjunto da disciplina Matemática, podendo ser Matemática I (aritmética), Matemática II (álgebra) e Matemática III (tópicos especiais). O desempenho nessas disciplinas, ao fim do ano, designa se o aluno foi aprovado ou não. Por isso todas são consideradas, como um grupo, para previsão do desempenho acadêmico da disciplina Matemática. Não obstante, têm-se alguns projetos e avaliações que podem conferir pontuação para disciplina Matemática como um todo, e esses também são considerados. A Tabela 2 descreve quais disciplinas estão associadas as "disciplinas mães", em relação aos identificadores de rede (Id.R) e de Série (Id.S).

5 Analise Descritiva dos Dados

De acordo com Fávero e Belfiore (2017) a estatística descritiva é responsável por descrever e resumir as principais características observadas nos dados. A fim de expandir o entendimento do conjunto de dados obtido, separaram-se as variáveis quantitativas das qualitativas. As variáveis qualitativas no caso são identificadores do registro de nota do aluno. Ou seja, são responsáveis

Id.R	Id.S	Disciplina Mãe	Disciplina
2	3	Matemática	Matemática
		Português	Português
	21	Matemática	Matemática, Matemática 1, Matemática 2, Matemática 3
		Português	Português
3	3	Matemática	Matemática
		Português	Português
	21	Matemática	Matemática, Matemática 1, Matemática 2
		Português	Português

Tabela 2: Disciplinas por rede, série e "disciplina mãe". Fonte: Autor.

Figura 1: Histograma de quantidade de registros por aluno. Fonte: Autor .

de indicar a qual rede, serie, aluno, período letivo, disciplina e avaliação pertencem à nota registrada por linha do conjunto de dados. Estatísticas a respeito desses dados fornecem informações importantes para o experimento proposto, que enriquecem a análise dos resultados subsequentes.

A Tabela 3 apresenta contagens simples a respeito do conjunto de variáveis qualitativas. As colunas na tabela representam respectivamente: identificador da rede, identificador do aluno, identificador da série, ano escolar, disciplina, disciplina mãe, ordem da avaliação no período letivo e nome da avaliação. Evidencia-se que a quantidade de registros por rede é aproximadamente homogêneo, uma vez que o conjunto contem duas redes, a rede com o identificador "3" é mais frequente, porém com 52% dos registros totais. Evidentemente têm-se alguns alunos com cerca de 90 registros, no entanto, ao contrastar com a Figura 1 chega-se a conclusão que são minoria na base, e geralmente os alunos possuem cerca de 40 registros cada. O valor discrepante pode apontar para registros duplicados, para determinado identificador de aluno.

O conjunto apresenta dois "ids", identificadores, que fazem referência ao primeiro ano do

	Id.R	Id.A	Id.S	Ano	Discip.	Discip. Mãe	Ord.	Av.
unic.	2	5792	2	2	5	2	19	77
top	3	596320	21	2018	Por.	Mat.	11	PM1-M.
freq.	124014	94	197144	122732	99300	135696	16758	5821

Tabela 3: Descrição das variáveis qualitativas, no que diz respeito a quantidade de registros únicos (unic.), valor mais frequente (top) e quantidade de registros associada ao valor mais frequente (freq.) Fonte: Autor .

ensino médio (série:21) e ao oitavo ano do ensino fundamental (série:3), respectivamente. Ainda de acordo com Tabela 3 têm-se os identificadores de rede (Id.R), aluno (Id.A) e série (Id.S), em conjunto com colunas que identificam o ano letivo (Ano), a disciplina (Discip.), a hierárquica agregadora de disciplina (Discip. Mãe), O número que identifica aquela avaliação no ano letivo, que pode ser entendido como a ordem de aplicação da avaliação (Ord.), e o nome da avaliação (Av.). Observa-se que alunos do primeiro ano são mais representativos na base de dados, isso se deve ao fato de as turmas do ensino médio possuírem subdivisões para disciplina Matemática, além de incorporarem mais alunos em relação ao oitavo ano.

No que diz respeito aos atributos quantitativos, neste conjunto de dados eles são dois: Nota e Aprovado. O primeiro se trata de uma representação numérica do total de pontos obtido pelo aluno na respectiva avaliação. Como cada avaliação possui um peso, esse valor pode ser em alguns casos maior que 1 (como pode ser observado na Tabela 5); o segundo se trata de uma variável categórica que informa se o aluno foi ou não aprovado naquele período letivo.

A partir da observação e do contraste entre as variáveis é possível atingir uma maior compreensão a respeito da dificuldade de determinadas disciplinas, em turmas específicas (Tabela 4). Para o primeiro ano do ensino médio (série:21) a disciplina Matemática apresenta uma média de notas baixas (média de 0,24 a 0,37) para ambas as redes.

Por outro lado, para o oitavo ano do ensino fundamental, notas baixas são mais incomuns, portanto podem ser um atributo mais relevante para previsão de desempenho acadêmico cedo. No que diz respeito ao aspecto temporal, as duas populações se comportaram de forma semelhante nos dois anos letivos, obedecendo distribuições de notas similares.

Ao se observar mais especificamente o aspecto da aprovação ao fim do período letivo, destaca-se a característica desbalanceada esperada em um conjunto de dados de previsão de desempenho acadêmico, sobretudo quando se busca inferir a aprovação de alunos. A grande maioria dos alunos é aprovada, em todas as turmas. As turmas que mais reprovaram foram as do primeiro ano (serie:21), principalmente na disciplina Matemática, em ambos os anos. Essa afirmação é aferida atravéz da observação da média da coluna "Aprovado"presente na Tabela **??**

6 Metodologia

A previsão de desempenho acadêmico para este trabalho foi modelada como um problema de classificação de sequência, que pode ser definida da seguinte forma: dada uma sequência $S_{(a)} = \{v_1, v_2, ..., v_n\}$ de avaliações realizadas pelo aluno a, procura-se aprender a função, f, que infere a probabilidade de a ser aprovado ou não ao fim do período letivo.

Ano	Id.R	Id.S	Disciplina	Nota cont.	med.	desv.	min.	50%	max.
2018	2	3	Mat.	5472,0	0,714584	0,399549	0,0	0,9000	1,00
			Por.	5472,0	0,712297	0,397003	0,0	0,9000	1,00
		21	Mat.	35938,0	0,247738	0,281026	0,0	0,1575	1,00
			Por.	19026,0	0,489657	0,337171	0,0	0,5400	1,00
	3	3	Mat.	4672,0	0,723513	0,372998	0,0	0,9000	1,00
			Por.	4672,0	0,715504	0,368930	0,0	0,9000	1,00
		21	Mat.	26114,0	0,376864	0,298960	0,0	0,3400	1,00
			Por.	21366,0	0,497679	0,331926	0,0	0,5400	1,00
2019	2	3	Mat.	3469,0	0,652749	0,432657	0,0	0,9000	1,00
			Por.	3469,0	0,653024	0,431936	0,0	0,9000	1,00
		21	Mat.	23608,0	0,258492	0,293293	0,0	0,1350	1,00
			Por.	14528,0	0,535800	0,379312	0,0	0,6000	1,00
	3	3	Mat.	5312,0	0,733141	0,372363	0,0	0,9100	1,00
			Por.	5314,0	0,734136	0,371776	0,0	0,9100	1,00
		21	Mat.	31111,0	0,355573	0,298176	0,0	0,3000	1,13
			Por.	25453,0	0,508643	0,338577	0,0	0,6000	1,00

Tabela 4: Descrição do atributo “nota”. Fonte: Autor .

Com objetivo de validar a hipótese estruturada e levando-se em consideração o conjunto de dados apresentado, se organizou o seguinte experimento. Separou-se o conjunto de dados em subconjuntos que contenham apenas um período letivo, uma rede, um ano escolar (ou série) e uma disciplina. Dessa forma tem-se 16 subconjuntos de dados. Logo se torna possível comparar o desempenho dos modelos avaliados em classificar se um aluno será aprovado ou não em duas populações (as duas redes), em disciplinas com diferentes características de dificuldade (Matemática e Português, avaliadas no oitavo ano do ensino fundamental e no primeiro ano do ensino médio), e em anos letivos distintos. Assim proporcionando uma validade mais robusta para os achados do experimento.

Em trabalhos recentes também se questionou a utilização de modelos de aprendizado profundo para previsão de sequências, especialmente séries temporais (Elsayed et al., 2021). O argumento principal vai de encontro a ideia que modelos como o *Gradient Boosting*, quando treinados em um conjunto de dados modelado de forma adequada, apresentam desempenho superior aos modelos de aprendizado profundo. Nesse sentido, um modelo interessante para contrastar os resultados por si só seria o supracitado *Gradient Boosting*.

Por fim, também foi construído uma rede *feedforward* de aprendizado profundo baseado em uma arquitetura mais simples para averiguar se existem ganhos tangíveis com o aumento da complexidade trazida pelo modelo baseado na arquitetura *Transformer*, na tarefa proposta. Logo serão avaliados três modelos, em 16 subconjuntos de dados, sendo eles:

- *Transformer*;
- *Xgboost*;

Ano	Id.R	Id.S	Disciplina	Aprovado cont.	med.	desv.
2018	2	3	Mat.	5472	0,877193	0,328246
			Por.	5472	0,877193	0,328246
		21	Mat.	35938	0,781457	0,413264
			Por.	19026	0,854305	0,352810
	3	3	Mat.	4672	0,910959	0,284834
			Por.	4672	0,904110	0,294472
		21	Mat.	26114	0,851727	0,355377
			Por.	21366	0,925864	0,261999
2019	2	3	Mat.	3469	0,833958	0,372172
			Por.	3469	0,833958	0,372172
		21	Mat.	23608	0,814978	0,388324
			Por.	14528	0,868943	0,337475
	3	3	Mat.	5312	0,915663	0,277919
			Por.	5314	0,915694	0,277872
		21	Mat.	31111	0,848671	0,358375
			Por.	25453	0,913016	0,281817

Tabela 5: Descrição do Atributo “aprovação”. Fonte: Autor .

- *Weighted* (sendo essencialmente uma rede *feedforward*).

O **modelo Transformer** se trata de uma adaptação da arquitetura *Behaviour Sequence Transformer* (BST) apresentada por Q. Chen et al. (2019). Nesse primeiro momento só são considerados os atributos relacionados a sequência de avaliações e a respectiva sequência de notas, diferente do BST original que usa outros atributos além da sequência. Em segundo lugar, por utilizar as notas dos alunos em provas na sequência de entrada em conjunto com sua posição na sequência, para serem atualizados antes de alimentarem o bloco de *Multi-Head Attention*, que se utiliza do mecanismo de atenção proposto por Bahdanau et al. (2014) para aprender representações ponderadas de uma sequência de entrada. A configuração do modelo se encontra descrita na Tabela 6. Adicionalmente, o modelo utilizou um mecanismo de “parada antecipada” (*early stopping*) para interromper o treinamento caso a métrica de erro do conjunto de validação não reduzisse em dez épocas.

A arquitetura adaptada pode ser visualizada na Figura 2. O modelo recebe a sequência de avaliações dos alunos, incluindo o resultado se ele foi aprovado ou não no período. Primeiro as variáveis de entradas são transformadas em uma representação vetorial (*embedding)*, de baixa dimensionalidade. Para melhor capturar a relação entre a sequência de avaliações, a camada *transformer* é utilizada para aprender representações profundas a respeito das avaliações na sequência. Então, a saída da camada *transformer* é concatenada e passada para as camadas *feedforward* da rede, usadas para aprender as interações dos atributos gerados na etapa anterior. A função sigmoide é usada para gerar a saída final do modelo.

A fim de se inferir o desempenho acadêmico de uma determinado aluno, no caso representado como um rótulo, se ele será aprovado ou não ao fim do período letivo, o problema foi

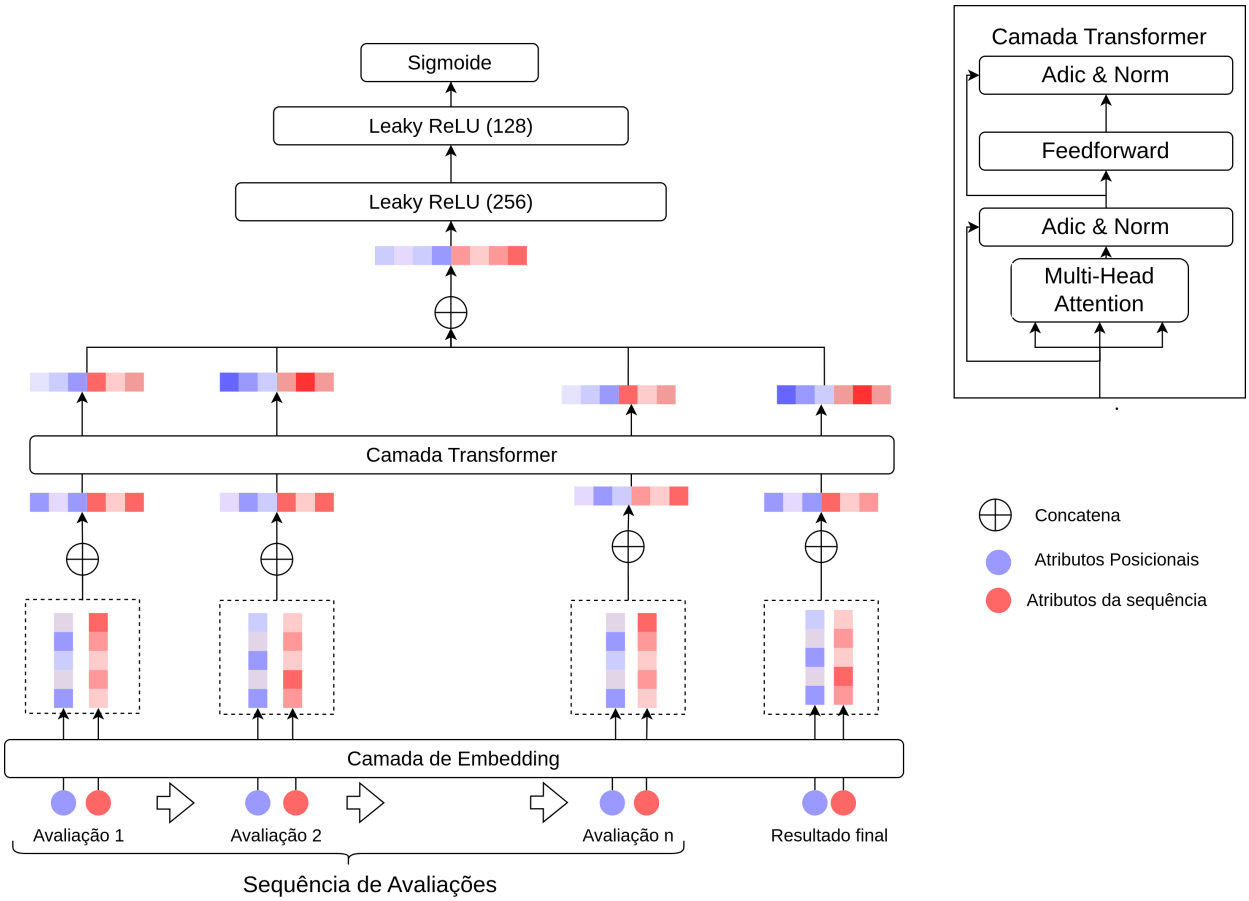

Figura 2: Visão geral da arquitetura trabalhada, baseada na BST proposta por Q. Chen et al. (2019). Fonte: Autor.

modelado como um problema de classificação binária, portanto foi utilizada a função sigmoide na última camada, tal qual proposto por Q. Chen et al. (2019), como paralelo no contexto de sistemas de recomendação. Para se treinar o modelo, foi utilizada a função de perda da entropia-cruzada (*cross-entropy loss*), definida pera Equação 1, onde D representa todas as amostras, e $y \in \{0,1\}$ o rótulo que indica se o aluno foi aprovado ou não ao fim do período, $p(x)$ é a saída da rede após o processamento pela unidade sigmoide, representando a probabilidade do aluno com a sequência x de atributos ser aprovado.

$$\lambda = -\frac{1}{N} \sum_{(x,y)\in D} (y \log p(x) + (1-y) \log(1-p(x))) \tag{1}$$

O **modelo Xgboost** lança mão da implementação apresentada em Chen e Guestrin (2016). Se trata de um modelo de rápido treinamento, com alta capacidade de generalização, e amplamente utilizado na literatura. Foram escolhidos 300 estimadores e uma taxa de aprendizado de 0,1; optou-se por se manter os outros parametros conforme o padrão fornecido pela biblioteca, e os valores escolhidos foram obtidos a partir de validação cruzada.

O **modelo Weighted** se trata de uma rede *feedforward* com três camadas ocultas da forma: [256 * 256 * 256], a função de ativação escolhida foi a unidade linear retificada (ReLu), e na

Configuração da Rede			
Tamanho de embedding	$1 \sim 16$	Tamanho do lote	64
Número de heads	8	Abandono	0,2
Tamanho da sequencia	$1 \sim 8$	Épocas	100
Bloco transformer	1	Taxa de aprendizado	0,01
Configuração MLP	256 * 128		

Tabela 6: Configuração do modelo baseado na arquitetura Transformer. Fonte: Autor .

camada de saída a sigmoide. Também foi utilizado o mecanismo de parada antecipada, da mesma forma que no modelo *transformer*, e no mesmo conjunto de épocas.

Como existe o interesse em entender a partir de quantas provas os modelos passam a apresentar desempenho interessante para aplicações que favorecem a identificação de alunos em risco de reprovação, cada modelo foi treinado em um subconjunto de dados contendo uma determinada quantidade de provas que varia de 1 até $n/2$, sendo n a quantidade total de avaliações. Isso porque se entende que a utilidade de modelos de previsão de desempenho acadêmico estão associadas a janela de atuação, conforme a janela de atuação das coordenações pedagógicas vai diminuindo se torna cada vez menos efetivo se realizar inferências a respeito do desempenho acadêmico dos alunos. Sendo assim, os modelos foram avaliados no que diz respeito a sua capacidade preditiva a partir de avaliações até a metade do período letivo, o que, no conjunto de dados observado, significa até um máximo de 6 a 8 avaliações.

Os conjuntos de dados foram separados em conjunto de treino e teste, sendo 20% reservado para teste e os outros 80% para treinamento. No caso dos modelos baseados em aprendizado profundo, 20% do conjunto de treino foi separado como conjunto de validação, utilizado para monitorar a métrica de perda durante o treinamento. Os resultados dispostos no trabalho são referentes a avaliação dos modelos no conjunto de teste. A separação dos conjuntos de dados foi feita de forma a manter a distribuição original dos dados, nos subconjuntos.

Apenas o modelo *Transformer* foi treinado para trabalhar com uma sequência de eventos sem nenhum pré-processamento. Nesse caso o modelo recebe como entrada um vetor com a sequência de avaliações e a sequência com os respectivos nomes de cada avaliação. O que resulta em um vetor x com apenas duas dimensões, onde a primeira é uma sequência de notas e a segunda uma sequência de nomes de avaliações referentes as notas. Um exemplo do conjunto de dados utilizado para o treinamento do modelo transformer pode ser observado na Tabela 7

IdAluno	NomeAvaliacao	NotaAvaliacao	Aprovado
1447	p1 - Matemática 1,p1 - Matemática 2	0,27,0,36	1
2054	p1 - Matemática 1,p1 - Matemática 2	0,2025,0,225	1
3874	p1 - Matemática 1,p1 - Matemática 2	0,315,0,405	1
4859	p1 - Matemática 1,p1 - Matemática 2	0,3825,0,36	1
6235	p1 - Matemática 1,p1 - Matemática 2	0,0,0,0	0

Tabela 7: Exemplo de entrada do modelo transformer. Fonte: Autor.

A fim de se treinar os outros modelos, o conjunto de dados foi pré-processado para que cada

dimensão do vetor x representasse a nota em uma avaliação. Para o conjunto de dados em questão, todos os alunos efetuaram todas as avaliações.

7 Material de Trabalho

Para se treinar os modelos de aprendizado profundo utilizou-se a biblioteca Keras (Chollet, 2015) pertencente a linguagem de programação Python (Oliphant, 2007). O Keras é uma biblioteca de aprendizado de máquina de alto nível projetada para a construção de modelos de aprendizado de máquina. Keras é uma biblioteca de abstração simples, que pode ser usada como uma camada sobre outras bibliotecas de aprendizado de máquina, como *TensorFlow*, *Theano* e CNTK. Ele fornece uma interface simplificada para construir modelos de aprendizado de máquina e possui uma vasta variedade de camadas, perda, otimizadores, modelos e outras funcionalidades para construir modelos de aprendizado de máquina. O Keras permite a construção de modelos de aprendizado de máquina complexos usando poucas linhas de código (Chollet, 2015). Ele também suporta modelos sequenciais (uma pilha linear de camadas) e modelos funcionais (permite criar modelos com fluxo de dados mais complexos) e também pode ser conectado ao TensorFlow, uma biblioteca para criação de modelos de aprendizado profundo, tornando-se possível acelerar o treinamento de redes neurais com GPUs (*Graphics Processing Unit*).

A fim de se treinar o modelo baseado em *Grandient Boosting*, utilizou-se a biblioteca *Xgboost* do Python (Chen & Guestrin, 2016). *XGBoost* é uma biblioteca de aprendizado de máquina de código aberto escrita em Python e implementada em C++. Ela foi criada para otimizar o desempenho do algoritmo *Grandient Boosting*, usando técnicas avançadas como regularização e distribuição paralela (Chen & Guestrin, 2016).

Para o treinamento dos modelos utilizaram-se máquinas aceleradas por GPU, por *Jupyter Notebooks* disponíveis pelo *Google Colaboratory*. O *Jupyter Notebook* é um aplicativo web de código aberto que permite criar e compartilhar documentos que contêm código, equações, visualizações e texto explicativo. A ferramenta segue o paradigma chamado de *literate programming* (programação letrada, em tradução livre). Ele é utilizado em diversas áreas, incluindo Ciência de Dados, Aprendizado de Máquina, Ensino de Ciência da Computação e Ciências. Os documentos criados no *Jupyter Notebook* são chamados de "*notebooks*" com a extensão "*.ipynb*". Um notebook consiste em uma série de células, que podem conter código, "*markdown*" (formato de texto) ou outro tipo de conteúdo, como imagens. As pessoas usuárias podem executar o código em cada célula e trabalhar com código de uma forma iterativa. Além disso, o *Jupyter Notebook* também possui recursos avançados, como suporte para visualização de dados, integração com bibliotecas de aprendizado de máquina populares e a capacidade de compartilhar os recursos criados com outras pessoas. *Google Colaboratory*, ou *Colab*, se trata um ambiente de desenvolvimento *Jupyter Notebook* gratuito criado pela Google. Ele é baseado no *Jupyter Notebook* e oferece acesso gratuito a GPUs (*Graphics Processing Unit*).

As máquinas aceleradas por GPU são computadores que possuem uma ou mais GPUs dedicadas para acelerar tarefas específicas, geralmente relacionadas a processamento gráfico ou computação paralela. Em comparação com as CPUs (*Central Processing Unit*) convencionais, as GPUs são projetadas para lidar com operações matemáticas complexas, como as necessárias para a renderização de gráficos tridimensionais, eficientemente (Goodfellow et al., 2016). As GPUs

conseguem realizar milhões de operações matemáticas simultaneamente, o que as torna ideias para tarefas de alta paralelismo como o treinamento de redes neurais. Diversos algoritmos de aprendizado de máquina, especialmente aqueles que envolvem grandes volumes de dados, podem ser acelerados significativamente ao serem executados em uma GPU. Uma revisão acerca dos aspectos transformadores das GPUs na ciência, com enfoque na indústria farmacêutica, é realizado por (Pandey et al., 2022).

8 Resultados

A métrica utilizada para comparação foi a área sob a curva de precisão e sensibilidade, justamente por se tratar de um problema desbalanceado em essência, e a princípio não ter sido feita nenhuma transformação no conjunto de dados para balanceá-los. Os resultados apresentando são referente ao desempenho dos modelos em relação aos conjuntos de teste.

Na Figura 3 é mostrado um gráfico de linha que resume o desempenho do modelo na métrica em questão, observando todos os 16 subconjuntos de dados em suas respectivas configurações. Ou seja, o eixo *x* representa o número de avaliações em que o modelo foi treinado no subconjunto dos dados e no eixo *y* seu respectivo desempenho agregado. Essa visualização é importante para contrastar não só o desempenho de cada modelo, mas a partir de quantas avaliações ele passa a ser útil naquele contexto.

Nota-se que o modelo *xgb* apresenta o melhor desempenho dos três na maioria das janelas de tempo, seguido pelo modelo *weighted* e por fim pelo *transformer*. A princípio o *transformer* possui o pior início – primeira avaliação –, dos três (0,8764 com uma prova, comparado a 0,9309 e 0,9416, do modelo *weighted* e *xgb* respectivamente), porém, a medida que vai se aumentado a quantidade de informação ingerida pelo modelo, mais especificamente a partir da sexta avaliação, o *transformer* conseguiu alcançar o patamar dos demais modelos (0,9560, 0,9497, 0,9568 - *transformer*, *weighted* e *xgb* respectivamente), e enfim superá-los em média (0,9740, 0,9436, 0,9410)

Separando-se as populações das duas redes, tem-se a Figura 4. Nela, percebe-se que todos os modelos tiveram mais dificuldade em generalizar os dados da rede 2, porém o *transformer* foi mais prejudicado. Isso pode ser percebido devido à curva com uma subida mais íngreme e longa que os outros modelos. A partir da quinta avaliação todos se mantiveram em um patamar próximo, levando-se em consideração a rede 2, todavia com tendência de queda entre a quinta e a sétima avaliação. Ao chegar-se na última avaliação as linhas divergem e o modelo *transformer* é o único que passa a ter um ganho de desempenho. No caso da rede 3, os modelos se comportaram de forma linear, dentro do intevalo analisado, no que tange ao ganho de desempenho. Por fim, ao se chegar na prova final, em ambos os cenários os modelos *transformer* e *weighted* apresentaram resultados próximos nas duas populações, o observado não se repete para o *xgb* que apresenta um decréscimo de desempenho médio de 3 pontos (*transformer*: 0,9754, 0,9733 ; *weighted*: 0,9324, 0,9493; *xgb*: 0,9193, 0,9519, nas redes 2 e 3 respectivamente).

Ao realizar a divisão do conjunto com base no recorte da série, é necessário levar em consideração os pontos identificados durante a análise exploratória. A princípio a série:3 (que representa o oitavo ano) apresenta uma taxa de aprovados, em média, mais alta que o primeiro ano

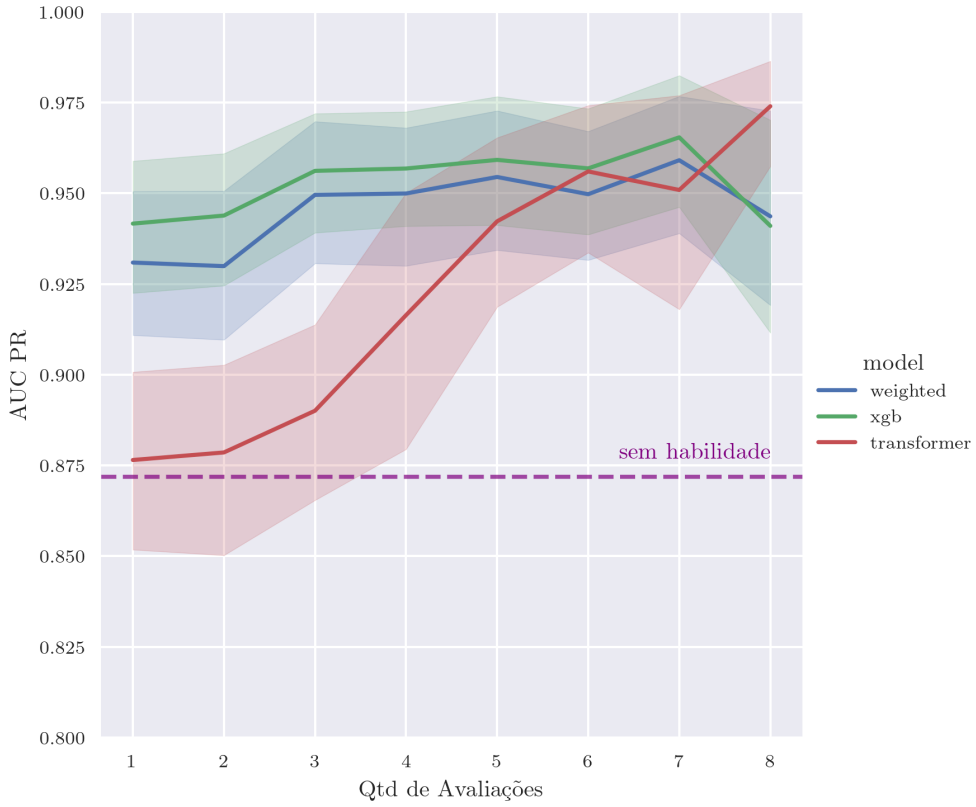

Figura 3: AUC PR por quantidade de avaliações por modelo, a linha pontilhada representa o classificador sem habilidade, que para uma curva de área sob a curva de precisão e sensibilidade representa a taxa de positivos no conjunto de dados avaliado. Fonte: Autor.

(série:21) e consequentemente é um problema mais desbalanceado. Por outro lado, parece ser mais simples para os modelos identificarem os casos de reprovação, isso pode ser notado por todos os modelos já começarem com um desempenho mais elevado na série:3 do que na série:21, com o *xgb* e o *weighted* seguindo a tendência de continuarem sendo relativamente superiores ao *transformer* para um conjunto pequeno de informação.

O ganho de informação observado pelo modelo *transformer* na série:21 é significativo (0,8555 para 0,9740, ao fim da sequência) fazendo com que o modelo inclusive supere os outros dois. Já para a série:3 o crescimento é um pouco menos expressivo (0,8974 para 0,9603) e levando os resultados dos modelos a um patamar parecido, com o *xgb* que chegou a 0,9856. Esse comportamento, de crescimento de desempenho alto, visto no *transformer* não é observado de forma tão latente nos outros dois modelos, que apresentam ganhos mais moderados.

Existe uma correlação alta do desempenho dos modelos ao se fazer o recorte de ano (*transformer*: 0,8855; *xgb*: 0,7418; *weighted* seguindo a tendência de continuarem sendo relativamente superiores ao *transformer*: 0,6939 para correlação de Pearson). Isto corrobora com o argumento de que as populações são inclusive comparáveis ano a ano (Fig. 6).

No que diz respeito as disciplinas, matemática é a disciplina com maior ganho de informação para o modelo *transformer* (0,8591 para 0,9809). De certa forma esse comportamento era esperado devido à disciplina matemática se subdividir em mais subconjuntos como mostrado anteriormente. Nesse caso, se espera que uma quantidade maior de provas e um entendimento mais

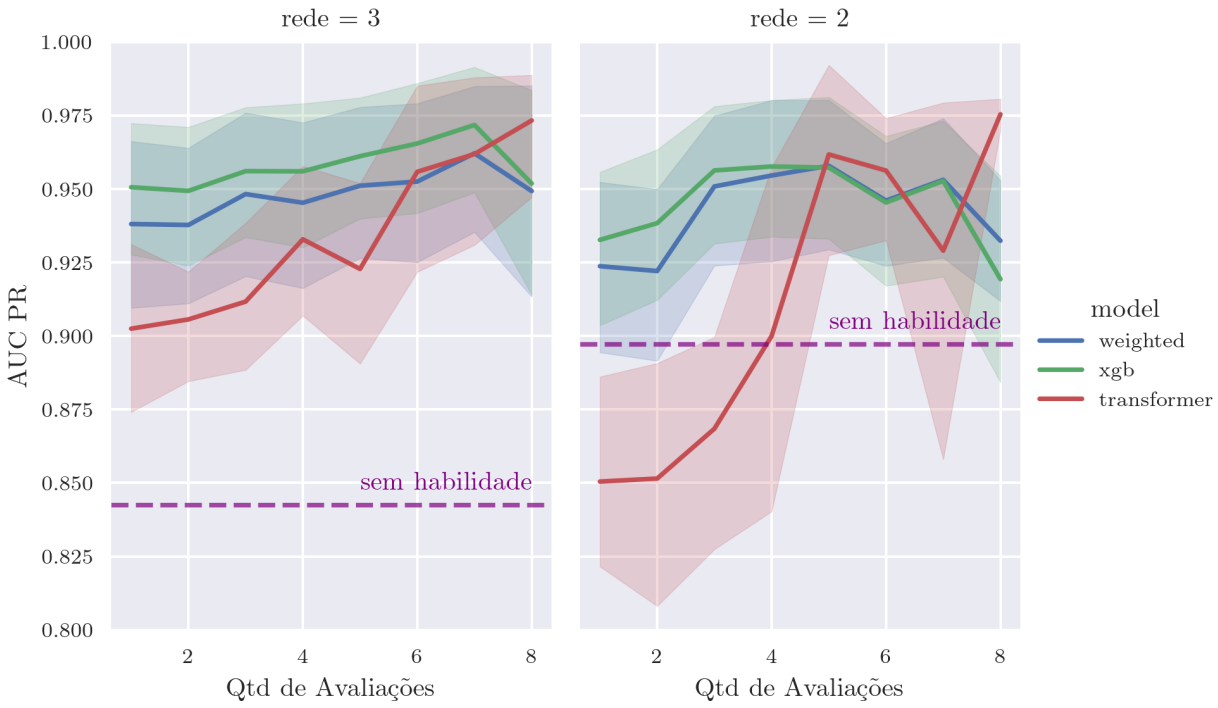

Figura 4: AUC PR por quantidade de avaliações por rede. Fonte: Autor.

model	Treinamento média	std	min	25 percentil	50 percentil	75 percentil	max
transformer	23,328571	6,968713	8,900	18,675	22,550	27,20000	52,6
weighted	8,130357	3,529077	2,600	5,800	7,600	11,00000	21,3
xgb	0,141920	0,168206	0,052	0,076	0,111	0,16875	1,8

Tabela 8: Estatísticas a respeito do tempo de treinamento dos modelos em segundos. Fonte: Autor.

representativo a respeito da sequência confira ao modelo uma maior capacidade de inferência a respeito daquele conjunto de dados. O mesmo efeito não se repete ao se observar o conjunto de português, justamente por esse não apresentar uma sequência mais extensa, consequentemente mais informação para o modelo *transformer*, ele apresenta ganho de desempenho, mas sem nunca cruzar o limite dos outros modelos.

Ao observa-se o tempo de treinamento dos modelos em segundos (Tabela 8) identifica-se que o tempo de treinamento do modelo *transformer* é o mais alto (em média) entre os modelos comparados. Isso se deve a complexidade envolvida no treinamento de um modelo baseado na arquitetura *Transformer*, e com a quantidade de parâmetros envolvidos, no entanto, não se trata de um tempo proibitivo. Por outro lado, o tempo registrado pelo modelo *xgb* é bem mais interessante do ponto de vista operacional do modelo.

Outro aspecto importante é a avaliação do ponto onde o desempenho do modelo supera o de um classificador sem habilidade, no caso da métrica utilizada esse ponto é representado pela média da taxa de positivos na população avaliada. Nesse caso, identificou-se esse ponto, para

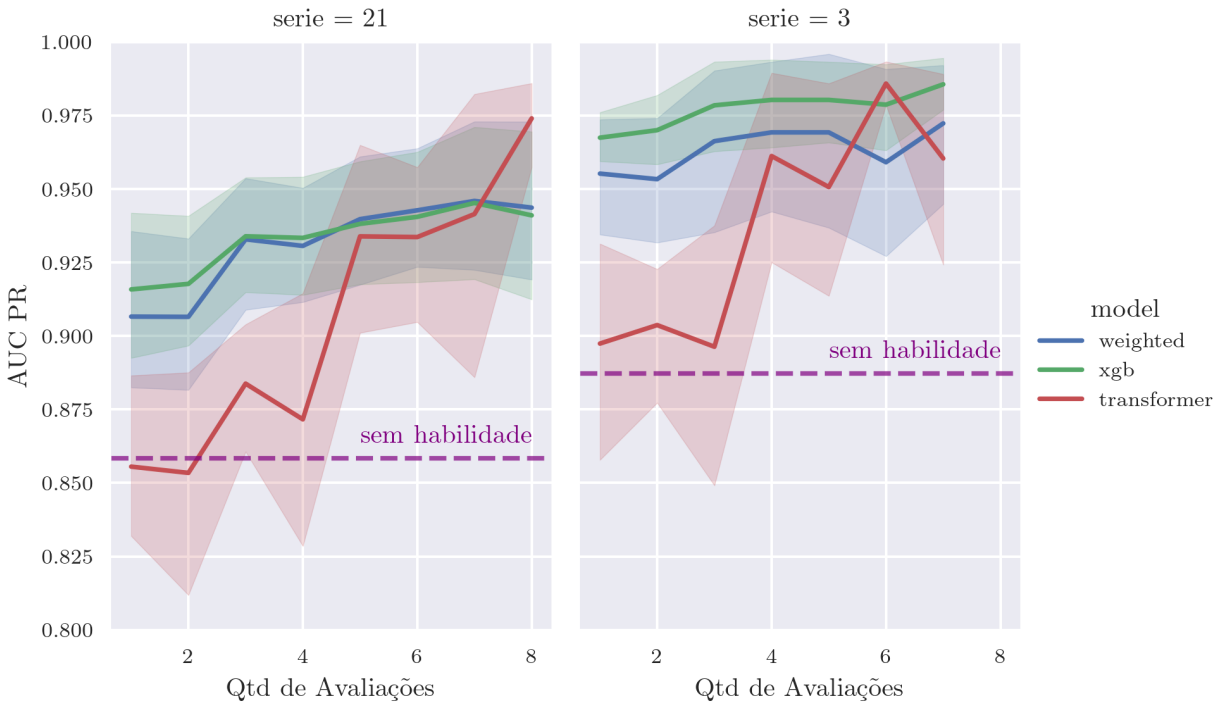

Figura 5: AUC PR por quantidade de avaliações por série. Fonte: Autor.

	Qtd de avaliações						
modelo	méd.	desv.	min.	25 percentil	50 percentil	75 percentil	max.
transformer	1,437500	0,629153	1	1	1	2	3
weighted	1,533333	1,597617	1	1	1	1	7
xgb	1,000000	0,000000	1	1	1	1	1

Tabela 9: Estatísticas a respeito do número de avaliações necessárias para os modelos ultrapassarem o limiar de um classificador sem habilidade .

cada conjunto de dados e construiu-se a Tabela 9. Nela nota-se que o modelo *transformer* possui resultados melhores que um classificador sem habilidade a partir da segunda leva de avaliações. O *xgb* apresenta o melhor desempenho, sendo um classificador hábil desde a primeira avaliação, diferente do modelo *weighted* que passa a responder de forma mais consistente entre a segunda/-terceira avaliação, levando-se em consideração o desvio padrão alto.

9 Discussão

Nesta seção será trabalhada a discussão das questões de pesquisa que rondam o trabalho de previsão de desempenho acadêmico a partir de aprendizado profundo para alunos do ensino fundamental e médio. Também serão abordadas as ameaças, a validade do estudo, e possíveis direções para trabalhos futuros.

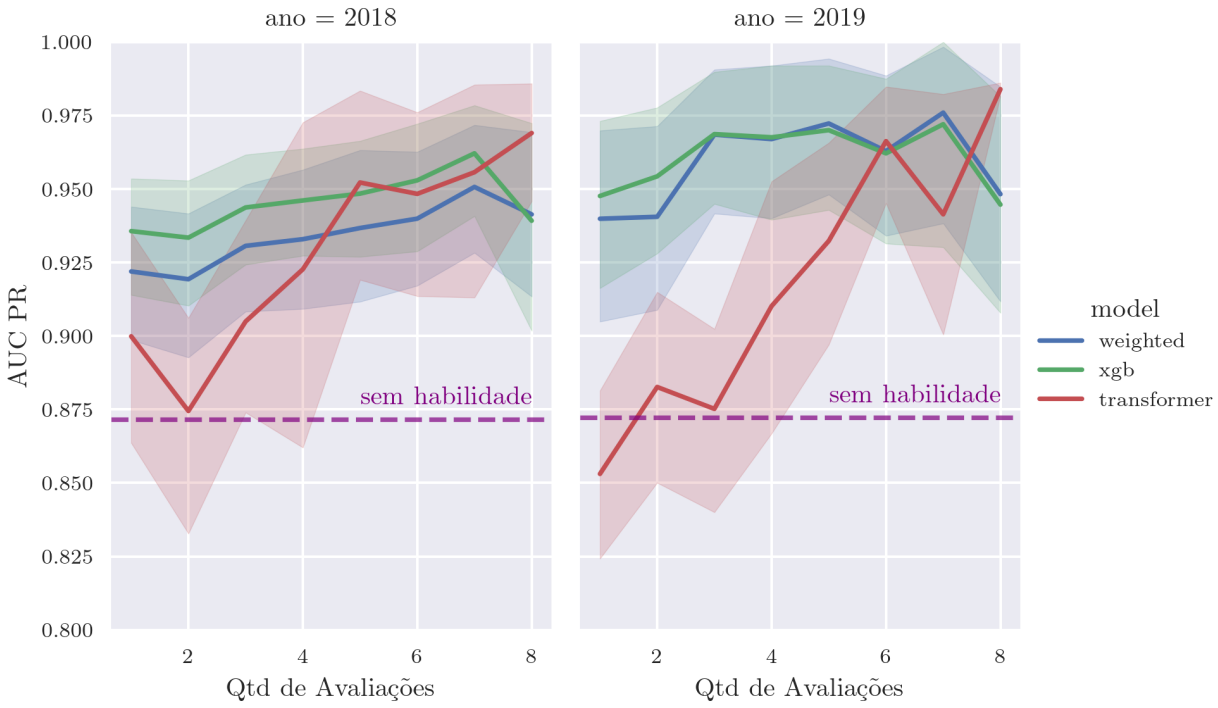

Figura 6: AUC PR por quantidade de avaliações por ano. Fonte: Autor.

9.1 Questão 1: Existe benefício em se utilizar aprendizado profundo, baseado na arquitetura *Transformer*, para previsão de desempenho acadêmico?

O modelo baseado na arquitetura *Transformer*, conforme apresentado na Seção 8, apresenta vantagens e desvantagens em relação aos modelos de comparação. A primeira desvantagem se trata do início mais lento, no que diz respeito a capacidade de inferência, ou seja, o modelo necessita de uma quantidade maior de avaliações para começar a apresentar um desempenho comparável com o *xgboost* ou a rede *feedforward* proposta. Porém, a partir de um número relativamente pequeno de avaliações (cerca de 6, de um total de, em média, 16 avaliações) o modelo chega ao patamar de desempenho dos outros, chegando eventualmente a superá-los.

Sugere-se que a razão para o modelo baseado na arquitetura *Transformer* apresentar um início mais "lento" a partir de um cenário com mais avaliações na sequência de entrada, e a medida que se tem acesso a mais informações aventar um desempenho melhor que os outros modelos, está relacionada a dois fatores.

O início mais "lento" pode ser explicado pelo modelo *transformer* ser um modelo que conhecidamente possui um desempenho melhor a partir de uma quantidade grande de dados (Xu et al., 2021), que não é o caso observado no experimento ao se ter de entrada uma sequência de poucas provas, no caso extremo, uma apenas.

A curva acentuada de aprendizado pode estar relacionada a uma melhor representação da sequência das avaliações, dada as características do modelo *transformer*. Para os outros dois modelos, a característica da sequência, da ordem das avaliações em si, é perdida. Isto em razão dos dados serem transformados de tal forma que cada dimensão do vetor de entrada representa

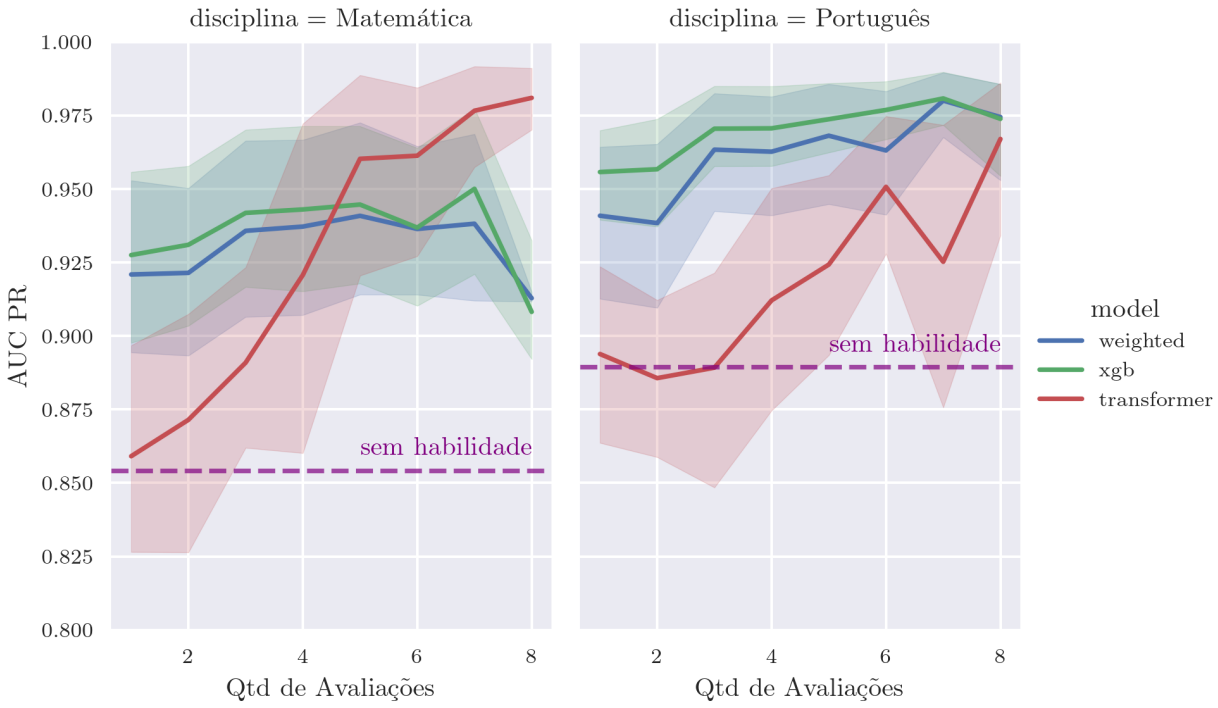

Figura 7: AUC PR por quantidade de avaliações por disciplina. Fonte: Autor.

uma avaliação, ou seja, os modelos não conhecem a ordem da sequência, sobretudo como elas se relacionam entre si. Assim, se conhece apenas os resultados das avaliações, e nesse caso o aspecto de evolução (ou de estagnação) por parte do aluno é perdido, caso essa informação não seja inserida por uma etapa de pré-processamento.

Por outro lado, a representação obtida pelo modelo *transformer* não perde a característica da sequência, inclusive por que tudo o que o modelo recebe na camada de entrada são as sequências de notas e avaliações, e o mecanismo de atenção é responsável por fornecer uma representação mais robusta dos dados que explicaria o desempenho maior dado um número grande de avaliações.

Um exemplo do comportamento de aumento do ganho de informação é observado na Fig. 7, em que para as avaliações da disciplina Matemática, que apresenta múltiplas subdivisões (Matemática 1, Matemática 2 e em alguns casos Matemática 3), o modelo *transformer* supera o desempenho dos outros modelos a partir da quinta avaliação e segue aumentando o desempenho de forma consistente. Esse comportamento não é observado com a disciplina de Português que não apresenta subdivisões, consequentemente menos avaliações para se aprender relações a respeito, e o modelo *transformer* não consegue atingir o patamar de desempenho dos outros dois. Essa constatação sugere a possibilidade de também cruzar-se a informação de múltiplas disciplinas para prever o resultado de uma específica, isso pode enriquecer a informação da sequência corroborando para um aumento da capacidade de inferência do modelo.

Existe um ganho também no que diz respeito à flexibilidade do modelo. Por se tratar de um modelo que não dispõem de pré-processamento, ele facilmente poderia ser aplicado a múltiplos cenários escolares sem se considerar os pormenores da transformação de dados, que tendem a ser problemáticos em sistemas produtivos. É necessário apenas considerar um *padding* do vetor

gerado na entrada – uma máscara para o vetor de entrada no treinamento para homogenizar a entrada da rede – suficientemente grande para não se ter problemas no treinamento.

Adicionalmente foi observada outra desvantagem, ao se observar o tempo de treinamento. Os modelos baseados na arquitetura *Transformer* possuem uma quantidade grande de parâmetros a serem treinados, impactando no tempo de treinamento. Diferente do *xgboost* treinado em média em menos de um segundo, nos conjuntos de dados propostos. Ainda considerando que o treinamento foi realizado em uma instância com GPU.

9.2 Questão 2: Caso haja benefício, a partir de qual ponto ele se mostra significativo, considerando o período letivo?

Segundo a Tabela 9, o modelo baseado na arquitetura *transformer* apresenta resultados significativos (superando um classificador sem habilidade), em média, a partir da segunda leva de avaliações. O que é um fator relevante para utilização do modelo em um sistema de aviso precoce, uma vez que esses sistemas visam identificar os alunos em risco de reprovação quanto antes, dando as coordenações pedagógicas tempo hábil para tratativa dos casos. As segundas avaliações, no caso das escolas presentes no conjunto de dados, ocorrem ainda no primeiro bimestre do ano, sendo positivo para aplicação do modelo.

O modelo baseado na arquitetura *transformer* não conseguiu superar o *xgboost* no que diz respeito a capacidade de inferência nas primeiras avaliações, visto que o segundo, de forma consistente, consegue superar o classificador sem habilidade a partir de uma avaliação. Porém, vale ressaltar que o modelo *transformer* supera a rede *feedforward* nesse quesito. Tanto no desempenho da tarefa em si, quanto na consistência da execução.

10 Ameaças à Validade

Existe um conjunto de ameaças a validade do estudo que serão elencadas nessa seção em conjunto com suas tratativas e possíveis impactos.

Observando-se o experimento proposto, tem-se que os modelos de aprendizado profundo são conhecidos pela sua natureza estocástica, o que prejudica em grande parte a reprodutibilidade de estudos. Nesse quesito, se almejou contornar essa dificuldade tomando-se múltiplos conjuntos de dados comparáveis entre si. Como relatado na Seção 4 tomou-se duas populações de alunos, em duas disciplinas, em anos letivos distintos, conferindo uma robustez aos resultados levantados. Outros aspecto é utilização de validação cruzada para o treinamento do modelo, uma vez que o conjunto de dados é tratado como uma sequência, a ordem possui um efeito importante nos testes. Uma melhoria nesse sentido seria a utilização de um método de validação cruzada que mantenha o aspecto ordenado do conjunto no treinamento, tal método não foi utilizado nesse experimento.

Em anos recentes, tivemos a pandemia de COVID-19 que teve seu efeito disruptivo na vida das pessoas e consequentemente nas escolas que na maior parte do Brasil foram fechadas e passaram a funcionar remotamente. Para não trazer uma característica tão desviante da norma, sobretudo na educação do ensino fundamental e médio, optou-se por utilizar conjuntos de dados que precedem ao período pandêmico. De certa forma, esta decisão permitiu a condução do presente

estudo, mas abre espaço para o levantamento de conjuntos de dados no período pós-pandêmico e o entendimento se os achados descritos neste trabalho se sustentam.

Os modelos propostos tiveram seus parâmetros ajustados, variando-os em faixas de valores, em dois subconjuntos de dados de treino com quatro avaliações cada, um subconjunto referente a cada rede. Ao fim escolheram-se os parâmetros utilizados nos modelos para o estudo na totalidade, com base nos modelos com melhor desempenho. No entanto, existe a possibilidade de aumento do escopo de busca dos melhores hiper parâmetros para cada período de avaliação. Ainda existe a possibilidade dos parâmetros encontrados não serem os mais otimizados devido ao limitando campo de busca, devido à limitações de hardware.

11 Conclusões Finais e Trabalhos Futuros

A presente trabalho teve como objetivo propor o uso de aprendizado profundo, mais especificamente uma adaptação da arquitetura *Transformer* para previsão de desempenho acadêmico de alunos do ensino fundamental e médio. Até o momento, entende-se que este é o primeiro trabalho a lançar mão da arquitetura por trás de modelos notáveis como o GPT e o BERT, tão bem sucedidos na área de processamento de linguagem natural, no contexto de previsão do desempenho escolar, visando o ensino básico.

Esse foco no ensino básico se deu justamente ao se observar que grande parte dos estudos se baseavam em populações de jovens do ensino superior ou associados a um curso *online* massivo. No entanto, é importante ressaltar, que os perfis de ensino superior e assinantes de cursos *online* massivos não refletem a realidade brasileira (majoritariamente pobre e pouco escolarizada). Assim sendo, a investigação de métodos que permitam a identificação precoce de alunos com risco reprovação nos cursos que são pilares da educação no país, tendem a ser benéficos para sociedade.

Logo, obteve-se um conjunto de dados de alunos do ensino básico, em duas disciplinas: Matemática e Português, onde fosse possível se contrastar anos escolares distintos, anos letivos distintos e populações distintas (redes de ensino diferentes). Em seguida, descreveu-se em detalhes o conjunto de dados obtido e o problema característico de previsão de desempenho acadêmico modelado. Tendo-se em vista a desconfiança da literatura mais recente em relação à capacidade de generalização dos modelos *transformers* em oposição a um modelo mais estabelecido como o *xgboost*, fez-se o esforço para contrastar os desempenho de ambos na tarefa de previsão de desempenho acadêmico, além de outra rede neural mais simples.

Os *transformers* apresentaram um desempenho interessante na tarefa proposta, principalmente a partir de um número maior de avaliações, contudo o *xgboost* consegue atingir um patamar elevado de acertos cedo no período letivo. Todavia, os *transformers* possuem uma flexibilidade grande no treinamento, desprezando pré-processamento, o que seria ideal para atuar em conjuntos de dados massivos de diversas instituições, por exemplo. No Brasil as escolas são obrigadas a apresentar o boletim, informando as avaliações e as notas dos alunos, porém não se tem um padrão de avaliações (rígido) o que pode dificultar o processamento de uma quantidade massiva de dados, por exemplo, para o *xgboost*. Por outro lado, o *transformers* pode se sair bem, uma vez que possui uma flexibilidade grande para trabalhar com sequências genéricas. Esse cenário se torna cada vez mais interessante quando se observa a constante digitalização dos conteúdos escolares,

consequentemente os próprios boletins.

Ademais, é de se imaginar também que modelos *transformers* treinados em conjuntos massivos (múltiplas escolas) com a adaptação proposta no presente trabalho, possuam uma grande capacidade de generalização – tal qual observado na evolução dos modelos de processamento de linguagem natural –, possivelmente capturando melhor as relações entre as sequências de avaliações e fornecendo melhores resultados. Quanto antes o sistema de aviso precoce consegue alertar as coordenações pedagógicas, mais rápido é possível trabalhar os casos críticos e consequentemente recuperar os alunos, e mitigar prejuízos a sociedade e a formação dos indivíduos.

Na perspectiva de desenvolvimentos futuros, identificam-se áreas promissoras para pesquisas adicionais. Uma delas é a exploração das potencialidades da transferência de aprendizado. Dada a eficácia da arquitetura baseada em *Transformers* em conjuntos extensos de dados pré-treinados, seria pertinente investigar o treinamento do modelo em diversas turmas e disciplinas. A avaliação subsequente do seu desempenho na previsão de uma nova turma pode seguir a abordagem adotada em Kim et al. (2018, 2019). Outra vertente de interesse é a pesquisa sobre o uso de informações provenientes de múltiplas disciplinas para inferir o desempenho em uma disciplina específica, ou seja, investigar o caráter de contribuição que uma sequência possui na inferência de outar. Um estudo mais aprofundado da otimização de hiperparâmetros do modelo proposto também se mostra relevante. Adicionalmente, seria valioso investigar o impacto da inclusão de mais informações sobre o aluno, como frequência, desempenho em anos anteriores, entre outros atributos considerados relevantes. Essa abordagem visa enriquecer a previsão do desempenho acadêmico, considerando uma gama mais ampla de variáveis.

Referências

Almayan, H., & Al Mayyan, W. (2016). Improving accuracy of students final grade prediction model using PSO (2ª ed.) [GS Search]. *Decision Analytics*, 35–39. https://doi.org/10.1109/INFOCOMAN.2016.7784211

Amra, I., & Maghari, A. (2017). Students performance prediction using KNN and Naïve Bayesian (8ª ed.) [GS Search]. *International Conference on Information Technology (ICIT)*, 909–913. https://doi.org/10.1109/ICITECH.2017.8079967

Athani, S. S., Kodli, S. A., Banavasi, M. N., & Hiremath, P. G. S. (2017). Student academic performance and social behavior predictor using data mining techniques (s.n) [GS Search]. *2017 International Conference on Computing, Communication and Automation (ICCCA)*, 170–174. https://doi.org/10.1109/CCAA.2017.8229794

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate [GS Search]. https://doi.org/10.48550/ARXIV.1409.0473

Barros, T. M., Souza Neto, P. A., Silva, I., & Guedes, L. A. (2019). Predictive Models for Imbalanced Data: A School Dropout Perspective [GS Search]. *Education Sciences*, *9*(44), 275. https://doi.org/10.3390/educsci9040275

Blasi, A. (2017). Performance increment of high school students using ANN model and sa algorithm [GS Search]. *Journal of Theoretical and Applied Information Technology*, *95*(11), 2417–2425.

Bonaccorso, G. (2017). *Machine learning algorithms* [GS Search]. Packt Publishing Ltd.

Chen & Guestrin, C. (2016, março). XGBoost: A Scalable Tree Boosting System [GS Search]. https://doi.org/10.1145/2939672.2939785

Chen, Q., Zhao, H., Li, W., Huang, P., & Ou, W. (2019). Behavior Sequence Transformer for E-Commerce Recommendation in Alibaba [GS Search]. *Proceedings of the 1st International Workshop on Deep Learning Practice for High-Dimensional Sparse Data*. https://doi.org/10.1145/3326937.3341261

Chen, W., Brinton, C. G., Cao, D., Mason-Singh, A., Lu, C., & Chiang, M. (2019). Early Detection Prediction of Learning Outcomes in Online Short-Courses via Learning Behaviors [GS Search]. *IEEE Transactions on Learning Technologies*, *12*(1), 44–58. https://doi.org/10.1109/TLT.2018.2793193

Chollet, F. e. a. (2015). Keras [GS Search].

Cornell-Farrow, S., & Garrard, R. (2020). Machine learning classifiers do not improve the prediction of academic risk: Evidence from Australia [GS Search]. *Communications in Statistics Case Studies Data Analysis and Applications*, *6*(2), 228–246. https://doi.org/10.1080/23737484.2020.1752849

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance (s.n) [GS Search]. *Proceedings of 5th Annual Future Business Technology Conference*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [GS Search]. https://doi.org/10.48550/ARXIV.1810.04805

Elsayed, S., Thyssens, D., Rashed, A., Schmidt-Thieme, L., & Jomaa, H. S. (2021). Do We Really Need Deep Learning Models for Time Series Forecasting? [GS Search]. *CoRR*, *abs/2101.02118*. https://arxiv.org/abs/2101.02118

Fávero, L. P., & Belfiore, P. (2017). *Manual de análise de dados: estatística e modelagem multivariada com Excel, SPSS e Stata* [GS Search]. Elsevier Brasil.

Fernandes, E., Holanda, M., Victorino, M., Borges, V., Carvalho, R., & Erven, G. (2019). Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil [GS Search]. *Journal of Business Research*, *94*, 335–343. https://doi.org/10.1016/j.jbusres.2018.02.012

García-González, J., & Skrita, A. (2019). Predicting academic performance based on students family environment: Evidence for Colombia using classification trees [GS Search]. *Psychology, Society and Education*, *11*(3), 299–311. https://doi.org/10.25115/psye.v11i3.2056

Gil, J., Delima, A., & Vilchez, R. (2020). Predicting students dropout indicators in public school using data mining approaches [GS Search]. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(1), 774–778. https://doi.org/10.30534/ijatcse/2020/110912020

Goldschmidt, R., Passos, E., & Bezerra, E. (2015). *Data mining* [GS Search]. Elsevier Brasil.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* [GS Search]. MIT press.

Gottardo, E., Kaestner, C. A. A., & Noronha, R. V. (2014). Estimativa de Desempenho Acadêmico de Estudantes: Análise da Aplicação de Técnicas de Mineração de Dados em Cursos a Distância [GS Search]. *Revista Brasileira de Informática na Educacação*, *22*(1). https://doi.org/10.5753/rbie.2014.22.01.45

Guo, B., Zhang, R., Xu, G., Shi, C., & Yang, L. (2016). Predicting Students Performance in Educational Data Mining (s.n) [GS Search]. *2015 International Symposium on Educational Technology (ISET)*, 125–128. https://doi.org/10.1109/ISET.2015.33

Guyon, I. (1991). Neural networks and applications tutorial [GS Search]. *Physics Reports*, *207*(3-5), 215–259.

H. Alamri, L., S. Almuslim, R., S. Alotibi, M., K. Alkadi, D., Ullah Khan, I., & Aslam, N. (2020). Predicting Student Academic Performance using Support Vector Machine and Random Forest (3ª ed.) [GS Search]. *2020 3rd International Conference on Education Technology Management*, (s.n), 100–107. https://doi.org/10.1145/3446590.3446607

Hellas, A., Ihantola, P., Petersen, A., Ajanovski, V. V., Gutica, M., Hynninen, T., Knutas, A., Leinonen, J., Messom, C., & Liao, S. N. (2018). Predicting academic performance: a systematic literature review (23ª ed.) [GS Search]. *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education*, (s.n), 175–199. https://doi.org/10.1145/3293881.3295783

Hernández-Blanco, A., Herrera-Flores, B., Tomás, D., & Navarro-Colorado, B. (2019). A Systematic Review of Deep Learning Approaches to Educational Data Mining [GS Search]. *Complexity*, *2019*, e1306039. https://doi.org/10.1155/2019/1306039

Hussain, S., Muhsin, Z., Salal, Y., Theodorou, P., Kurtolu, F., & Hazarika, G. (2019). Prediction model on student performance based on internal assessment using deep learning [GS Search]. *International Journal of Emerging Technologies in Learning*, *14*(8), 4–22. https://doi.org/10.3991/ijet.v14i08.10001

Hussain, S., & Khan, M. Q. (2021). Student-Performulator: Predicting Students Academic Performance at Secondary and Intermediate Level Using Machine Learning [GS Search]. *Annals of Data Science*, *s.n*, 1–19.

Imran, M., Latif, S., Mehmood, D., & Shah, M. (2019). Student academic performance prediction using supervised learning techniques [GS Search]. *International Journal of Emerging Technologies in Learning*, *14*(14), 92–104. https://doi.org/10.3991/ijet.v14i14.10310

Kim, B., Vizitei, E., & Ganapathi, V. (2018). GritNet: Student Performance Prediction with Deep Learning [GS Search]. *CoRR*, *abs/1804.07405*. http://arxiv.org/abs/1804.07405

Kim, B.-H., Vizitei, E., & Ganapathi, V. (2019). Domain Adaptation for Real-Time Student Performance Prediction [GS Search]. https://doi.org/10.48550/arXiv.1809.06686

Lee, S., & Chung, J. Y. (2019). The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction [GS Search]. *Applied Sciences*, *9*(1515), 3093. https://doi.org/10.3390/app9153093

Livingstone, D. J. (2008). *Artificial neural networks: methods and applications* [GS Search]. Springer.

Liz-Domínguez, M., Caeiro-Rodríguez, M., Llamas-Nistal, M., & Mikic-Fonte, F. A. (2019). Systematic Literature Review of Predictive Analysis Tools in Higher Education [GS Search]. *Applied Sciences*, *9*(2424), 5569. https://doi.org/10.3390/app9245569

Lu, H., & Yuan, J. (2018). Student performance prediction model based on discriminative feature selection [GS Search]. *International Journal of Emerging Technologies in Learning*, *13*(10), 55–68. https://doi.org/10.3991/ijet.v13i10.9451

Márquez-Vera, C., Cano, A., Romero, C., & Ventura, S. (2013). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional

and imbalanced data [GS Search]. *Applied Intelligence*, *38*(3), 315–330. https://doi.org/10.1007/s10489-012-0374-8

Mitchell, T. M., et al. (2007). *Machine learning* (Vol. 1) [GS Search]. McGraw-hill New York.

Namoun, A., & Alshanqiti, A. (2021). Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review [GS Search]. *Applied Sciences*, *11*(11), 237. https://doi.org/10.3390/app11010237

Niu, Z., Zhong, G., & Yu, H. (2021). A review on the attention mechanism of deep learning [GS Search]. *Neurocomputing*, *452*, 48–62. https://doi.org/10.1016/j.neucom.2021.03.091

Oliphant, T. E. (2007). Python for Scientific Computing [GS Search]. *Computing in Science Engineering*, *9*(3), 10–20. https://doi.org/10.1109/MCSE.2007.58

Orooji, M., & Chen, J. (2019). Predicting Louisiana Public High School Dropout through Imbalanced Learning Techniques [GS Search]. *CoRR*, *abs/1910.13018*. http://arxiv.org/abs/1910.13018

Pandey, M., Fernandez, M., Gentile, F., Isayev, O., Tropsha, A., Stern, A. C., & Cherkasov, A. (2022). The transformational role of GPU computing and deep learning in drug discovery [GS Search]. *Nature Machine Intelligence*, *4*(3), 211–221. https://doi.org/10.1038/s42256-022-00463-x

Qazdar, A., Er-Raha, B., Cherkaoui, C., & Mammass, D. (2019). A machine learning algorithm framework for predicting students performance: A case study of baccalaureate students in Morocco [GS Search]. *Education and Information Technologies*, *24*(6), 3577–3589. https://doi.org/10.1007/s10639-019-09946-8

Razaque, A., & Alajlan, A. (2020). Supervised machine learning model-based approach for performance prediction of students [GS Search]. *Journal of Computer Science*, *16*(8), 1150–1162. https://doi.org/10.3844/jcssp.2020.1150.1162

Rodrigues, L. S., dos Santos, M., Costa, I., & Moreira, M. A. L. (2022). Student Performance Prediction on Primary and Secondary Schools-A Systematic Literature Review [GS Search]. *Procedia Computer Science*, *214*, 680–687. https://doi.org/10.1016/j.procs.2022.11.229

Romero, C., Ventura, S., Pechenizkiy, M., & Baker, R. S. (2010). *Handbook of educational data mining* [GS Search]. CRC press.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. [GS Search]. *Psychological review*, *65*(6), 386. https://doi.org/10.1037/h0042519

Rovira, S., Puertas, E., & Igual, L. (2017). Data-driven system to predict academic grades and dropout [GS Search]. *PLOS ONE*, *12*(2), 1–21. https://doi.org/10.1371/journal.pone.0171207

Roy, S., & Garg, A. (2017). Predicting academic performance of student using classification techniques (4ª ed.) [GS Search]. *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, 568–572. https://doi.org/10.1109/UPCON.2017.8251112

Sharma, S., Sharma, S., & Athaiya, A. (2017). Activation functions in neural networks [GS Search]. *towards data science*, *6*(12), 310–316.

Sokkhey, P., & Okazaki, T. (2020). Study on dominant factor for academic performance prediction using feature selection methods [GS Search]. *International Journal of Advanced Computer Science and Applications*, *11*(8), 492–502. https://doi.org/10.14569/IJACSA.2020.0110862

Souza, V. F. d., & Santos, T. C. B. d. (2021). Processo de Mineração de Dados Educacionais aplicado na Previsão do Desempenho de Alunos: Uma comparação entre as Técnicas de Aprendizagem de Máquina e Aprendizagem Profunda [GS Search]. *Revista Brasileira de Informática na Educação*, *29*, 519–546. https://doi.org/10.5753/rbie.2021.29.0.519

Tatar, A. E., & Dütegör, D. (2020). Prediction of Academic Performance at Undergraduate Graduation: Course Grades or Grade Point Average? [GS Search]. *Applied Sciences*, *10*(1414), 4967. https://doi.org/10.3390/app10144967

Turabieh, H. (2019). Hybrid machine learning classifiers to predict student performance (2ª ed.) [GS Search]. *2019 2nd International Conference on new Trends in Computing Sciences (ICTCS)*, 1–6. https://doi.org/10.1109/ICTCS.2019.8923093

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need [GS Search]. https://doi.org/10.48550/ARXIV.1706.03762

Xiao, J., & Zhou, Z. (2020). Research progress of RNN language model [GS Search]. *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, 1285–1288. https://doi.org/10.1109/ICAICA50127.2020.9182390

Xu, P., Kumar, D., Yang, W., Zi, W., Tang, K., Huang, C., Cheung, J. C. K., Prince, S. J., & Cao, Y. (2021). Optimizing Deeper Transformers on Small Datasets [GS Search]. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2089–2102. https://doi.org/10.18653/v1/2021.acl-long.163

Yang, F., & Li, F. (2018). Study on student performance estimation, student progress analysis, and student potential prediction based on data mining [GS Search]. *Computers and Education*, *123*, 97–108. https://doi.org/10.1016/j.compedu.2018.04.006

Zaffar, M., Hashmani, M., Savita, K., & Rizvi, S. (2018). A study of feature selection algorithms for predicting students academic performance [GS Search]. *International Journal of Advanced Computer Science and Applications*, *9*(5), 541–549. https://doi.org/10.14569/IJACSA.2018.090569

Zhang, L., & Li, K. F. (2018). Education Analytics: Challenges and Approaches (1ª ed.) [GS Search]. *2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, 193–198. https://doi.org/10.1109/WAINA.2018.00086