UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

PAULO VITOR SOUTO DANTAS

CLASSIFICAÇÃO AUTOMÁTICA DE QUESTÕES CONFORME A TAXONOMIA DE BLOOM

CAMPINA GRANDE - PB

2021

PAULO VITOR SOUTO DANTAS

CLASSIFICAÇÃO AUTOMÁTICA DE QUESTÕES CONFORME A TAXONOMIA DE BLOOM

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Professor Dr. Cláudio Elízio Calazans Campelo.

CAMPINA GRANDE - PB

2021

D192c Dantas, Paulo Vitor Souto.
Classificação automática de questões conforme a Taxonomia de Bloom. / Paulo Vitor Souto Dantas. – 2021.

10 f.

Orientador: Prof. Dr. Cláudio Elízio Calazans Campelo.
Trabalho de Conclusão de Curso - Artigo (Curso de Bacharelado em Ciência da Computação) - Universidade Federal de Campina Grande; Centro de Engenharia Elétrica e Informática.

1. Taxonomia de Bllom. 2. Classificação automática de questões. 3. Aumento de base de dados. 4. Algoritmo XGNoost. 5. Algoritmo CatBoost. 6. Algoritmo SVM. 7. Algoritmo Random Forest. I. Campelo, Cláudio Elízio Calazans. II. Título.

CDU:004.912(045)

Elaboração da Ficha Catalográfica:

Johnny Rodrigues Barbosa
Bibliotecário-Documentalista
CRB-15/626

PAULO VITOR SOUTO DANTAS

CLASSIFICAÇÃO AUTOMÁTICA DE QUESTÕES CONFORME A TAXONOMIA DE BLOOM

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Professor Dr. Cláudio Elízio Calazans Campelo
Orientador – UASC/CEEI/UFCG

Professor Dr. Eanes Torres Pereira
Examinador – UASC/CEEI/UFCG

Professor Dr. Tiago Lima Massoni
Professor da Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 25 de maio de 2021.

CAMPINA GRANDE – PB

ABSTRACT

The diversification of the difficulty in the questions proposed to students has a direct impact on their learning process. Thus, the fundamental role of the teacher is to provide questions that encourage the critical sense of students to improve their cognitive skills. Therefore, a taxonomy that aims to assist this process, classifying the cognitive level required by the questions, is Bloom's taxonomy. In this context, computing can offer tools for the automatic classification of questions according to Bloom's taxonomy, benefiting teachers and students. Although there are works with the objective of creating automatic classifiers for Bloom's taxonomy, some algorithms that use Gradient Boosting techniques are not commonly used for this process. Therefore, this work proposes the use of the XGBoost and CatBoost algorithms to be compared with the SVM and Random Forest algorithms in the question classification process. In addition, we propose the use of automatic techniques to increase the number of questions, classified according to Bloom's taxonomy, available in the database. We believe that the result of this work contributes to the improvement of the question classification models.

Classificação automática de questões conforme a taxonomia de Bloom

Trabalho de Conclusão de Curso

Paulo Vitor Souto Dantas (Aluno), Cláudio Campelo (Orientador)
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba - Brasil

RESUMO

A diversificação de dificuldade das questões propostas para alunos tem um impacto direto no seu processo de aprendizagem. Assim, o papel fundamental do professor é fornecer questões que provoquem o senso crítico dos alunos para o aprimoramento de suas habilidades cognitivas. Portanto, uma taxonomia que visa auxiliar nesse processo, classificando o nível cognitivo exigido pelas questões, é a taxonomia de Bloom. Nesse contexto, a computação pode oferecer ferramentas para a classificação automática de questões de acordo com a taxonomia de Bloom, beneficiando professores e alunos. Embora existam trabalhos com o objetivo de criar classificadores automáticos para a taxonomia de Bloom, alguns algoritmos que utilizam técnicas de *Gradient Boosting* não são comumente utilizados para este processo. Portanto, este trabalho propõe a utilização dos algoritmos *XGBoost* e *CatBoost* para serem comparados com os algoritmos *SVM* e *Random Forest* no processo de classificação de questões. Além disso, propomos o uso de técnicas automáticas para aumentar o número de questões, classificadas de acordo com a taxonomia de Bloom, disponíveis na base de dados. Com isso, acreditamos que o resultado deste trabalho contribui para o aprimoramento dos modelos de classificação de questões.

PALAVRAS-CHAVE

Taxonomia de Bloom, Aumento de base de dados, XGBoost, CatBoost, Classificação automática de questões

1 INTRODUÇÃO

A aprendizagem é um processo que ocorre por meio da experiência. As pessoas aprendem a partir do ponto que se tornam capazes de realizar atividades diferentes, como afirma Schunk [7]. De acordo com Crowe e Wenderoth [2], a memorização e recordação requerem um mínimo nível de compreensão, enquanto a aplicação de conhecimento e o pensamento crítico são habilidades de ordem superior.

Alinhado a isso, Gardiner [5] retrata que existe uma relação entre a diversidade de cursos no currículo de um aluno e o ganho de habilidades intelectuais do mesmo. Sendo assim, diversificar os níveis dos desafios propostos para alunos seria uma maneira de influenciar nas habilidades do estudante. Deste modo, cabe ao professor aplicar um conjunto de questões que exijam diferentes níveis cognitivos dos alunos em exames acadêmicos. Para tal finalidade, uma ferramenta amplamente aceita é a taxonomia de Bloom de domínios cognitivos [3]. Com ela, é possível categorizar questões em seis classes, de acordo com o nível cognitivo exigido por elas: conhecimento, compreensão, aplicação, análise, síntese e avaliação.

Embora existam esforços para a criação de classificadores automáticos de questões, há uma escassez na utilização de algoritmos que utilizam-se de técnicas de *Boosting*, e de abordagens para aumento de base de dados, para verificar a influência no treinamento de modelos.

Diante disto, este artigo apresenta os resultados obtidos com a implementação de classificadores, utilizando diferentes algoritmos de aprendizagem de máquina supervisionada: *XGBoost* e *CatBoost* (ainda não avaliados por trabalhos anteriores); e *SVM* e *Random Forest* (já avaliados em trabalhos anteriores). Além disso, propõe-se a utilização de uma técnica de aumento de base de dados (*Data Augmentation*), que utiliza sinônimos para aumentar a quantidade de sentenças disponíveis para o processo de treinamento.

Para treino e teste dos classificadores, foram utilizadas três diferentes bases de dados: uma base de dados criada por Mohammed e Omar [6], contendo 141 questões rotuladas; uma base com 600 questões, também disponibilizada por Mohammed e Omar [6]; e uma base de dados proposta por Basu et al. [8] com 1284 questões.

Ao longo do trabalho, foram utilizadas estas técnicas para avaliar o impacto na rotulagem automática de questões. Assim, com o uso dos algoritmos citados, acreditamos que tais técnicas caracterizam uma boa estratégia para o aperfeiçoamento de modelos de classificação e possam auxiliar na escolha de abordagens para trabalhos futuros.

2 TRABALHOS RELACIONADOS

A aplicação de técnicas para a classificação automática de questões de acordo com a taxonomia de Bloom tem sido objeto de estudo recorrente nos últimos anos. Trabalhos utilizam diversas técnicas no treinamento de modelos para a classificação de tais questões, como, por exemplo, no trabalho de Mohammed e Omar [6], onde foram utilizadas técnicas de TF-IDF juntamente com POS Tagging para identificar a frequência dos termos de uma questão de acordo com sua classe gramatical, em conjunto com Word2vec pré-treinado para impulsionar o processo de classificação. Além disso, foram utilizados os algoritmos *K-Nearest Neighbour* (KNN), *Logistic Regression* e *Support Vector Machine (SVM)* para o treinamento dos

classificadores, e obtiveram 71,1%, 82,3% e 83,7% de F1, respectivamente.

Por sua vez, Basu et al. [8] fazem uso de duas abordagens com os algoritmos BERT e LDA, onde foram obtidas as acurácias de 89% e 83% respectivamente, e utilizaram como abordagem para aumento da base de dados a anotação manual, sem a utilização de nenhuma técnica automática para aumentar a base de treinamento. Por outro lado, Yahya et al. [11] utilizam-se do algoritmo *SVM* para o processo de treinamento dos modelos de classificação, e encontraram dificuldades devido a pouca quantidade de dados disponíveis para o treino.

Outra abordagem interessante pode ser observada no trabalho de Abduljabbar e Omar [1], onde os autores utilizam ensemble, com algoritmos de votação, para combinar três classificadores distintos. Com isso, obtiveram 92.28% como resultado de F1, sem utilizar nenhum dos algoritmos que utilizam técnicas de *Gradient boosting* para o processo de treinamento.

Diferente dos demais trabalhos, este propõe a utilização de uma técnica de *Data Augmentation* para maximizar a quantidade de questões na base de dados. Esta técnica se mostra promissora para esta problemática, tendo em vista a pouca quantidade de dados disponíveis, diferentemente de outros problemas de classificação textual que possuem maior disponibilidade de dados, possibilitando, inclusive, o uso de algoritmos de aprendizagem profunda. Além disso, foram utilizados dois classificadores treinados com os algoritmos *XGBoost* e *CatBoost*, para comparação com o *SVM* e *Random Forest*.

O restante deste artigo está organizado da seguinte forma: a próxima seção (Seção 3) tem como objetivo apresentar conceitos importantes para o entendimento de alguns assuntos abordados no trabalho, como: a Taxonomia de Bloom, o processo utilizado para o treinamento de modelos de classificação e a técnica de aumento de base de dados utilizada. Em seguida, a Seção 4 descreve a metodologia utilizada para o desenvolvimento dos experimentos com os modelos de classificação. Os resultados obtidos a partir da utilização desta metodologia, por sua vez, são discutidos na Seção 5. Por fim, a Seção 6 conclui o artigo e aponta para trabalhos futuros.

3 FUNDAMENTAÇÃO TEÓRICA

3.1 Taxonomia de Bloom

A taxonomia de Bloom [3], foi proposta em 1956 e se subdivide em 3 diferentes domínios para a classificação da aprendizagem: o cognitivo, afetivo e psicomotor. Neste trabalho, iremos abordar o domínio cognitivo, por ser responsável pela classificação dos objetivos educacionais. Tais objetivos são categorizados hierarquicamente em seis diferentes níveis de acordo com o nível cognitivo exigido por ele, são eles:

(1) **Conhecimento**: Habilidade cognitiva relacionada a retenção de conhecimento específico e discreto.
(2) **Compreensão**: É a capacidade de encontrar significado nas informações, parafrasear, agrupar informações e comparar com outras existentes.
(3) **Aplicação**: Capacidade de utilizar conhecimentos, habilidades ou técnicas em novas situações.
(4) **Análise**: Consiste no nível das habilidades de pensamento crítico. Neste nível o indivíduo deve ser capaz de distinguir entre fato e opinião e identificar os argumentos sobre os quais uma afirmação é construída.
(5) **Síntese**: Este nível corresponde a habilidade de criar um novo produto em uma determinada situação.
(6) **Avaliação**: Este é o nível mais alto da taxonomia e está associado ao julgamento sobre o valor de certa seção de um texto, avaliando ela criticamente.

Desta maneira, o primeiro nível (Conhecimento) necessita de menor habilidade cognitiva para ser alcançado, enquanto Avaliação é o nível que necessita de maior poder cognitivo. Assim, como está ilustrado na pirâmide da Figura 1, cada nível contém as habilidades exigidas pelos níveis inferiores a ele, além de novos desafios cognitivos.

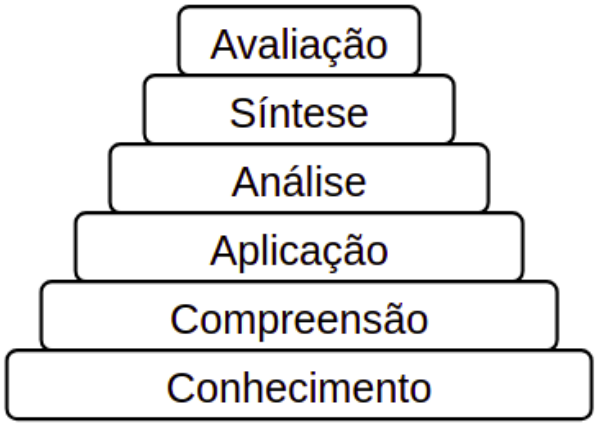

Figura 1: Pirâmide da taxonomia de Bloom.

3.2 Treinamento de modelos de classificação

Processamento de linguagem natural (*Natural Language Processing* - NLP) é uma área de pesquisa e aplicação que tem como objetivo utilizar computadores para manipulação de texto e fala [4]. Desta maneira, algoritmos são utilizados para efetuar o treinamento de modelos, que se tornarão capazes de entender a linguagem natural e chegar a determinadas conclusões a partir de uma observação. Normalmente, para problemas de classificação textual, são exploradas abordagens através do uso de algoritmos de aprendizagem de máquina supervisionados, onde se faz necessário a utilização de dados retirados do domínio de estudo e rotulados com a classe que o modelo deverá inferir. Assim, o algoritmo irá observar tais dados e aprender o padrão necessário para classificá-los. Com o reconhecimento destes padrões o modelo será capaz de atribuir uma classe (rótulo) a novas observações.

Para tal finalidade, os dados precisam passar por uma etapa de pré-processamento antes de serem utilizados pelo algoritmo de treinamento. Tal processo tem como objetivo a preparação e organização dos dados, e é iniciado com uma etapa de remoção de dados faltantes e remoção de dados irrelevantes, como, por exemplo: dados numéricos, simbólicos e *stop words*. Após esta etapa, as palavras contidas na base de dados passam por um processo de contração, que pode ser efetuado por meio da Lematização ou *Stemming*. Tal

processo é responsável por flexionar uma palavra para a sua forma básica.

Por fim, como última etapa do pré-processamento, é necessário utilizar técnicas para a extração de informações relevantes do texto. Deste modo, uma das técnicas mais utilizadas para tal procedimento é a conhecida como *TF-IDF*.

3.2.1 TF-IDF. Tem como objetivo indicar a importância de um termo dentro de um documento da base de dados, com base no número de vezes que ele aparece. Além disso, ele verifica a importância deste termo levando em consideração toda a base de dados disponível, ou seja, é responsável por indicar a raridade de um termo dentro dos documentos analisados. Para tal finalidade é utilizada a Equação 1:

$$w_{i,j} = tf_{i,j} \times log(\frac{N}{df_i}) \quad (1)$$

Sendo,

- $tf_{i,j}$: A quantidade de vezes que i aparece no documento.
- N: A quantidade total de documentos.
- df_i: A quantidade de documentos que contém i.

3.3 *Aumento de base de dados*

Técnicas de aumento de base de dados (*Data Augmentation*) tem como objetivo gerar novos dados a partir de dados já existentes. Assim, tais técnicas são utilizadas em bases de dados de treino para fornecer aos algoritmos de classificação uma visão diferente de um dado existente. A técnica que foi utilizada neste trabalho é baseada na substituição de sinônimos. Para efetuar tal procedimento, o algoritmo gera uma nova instância para uma frase a partir da substituição de uma palavra por um de seus sinônimos.

4 METODOLOGIA

Esta seção descreve a metodologia utilizada para a realização dos experimentos com os modelos de classificação, assim, são apresentados os seguintes tópicos: a abordagem utilizada para coletar os dados; as técnicas de pré-processamento utilizadas nas questões; a descrição do processo de treino e teste dos classificadores; o aumento da base de dados; e uma breve descrição sobre as métricas utilizadas para efetuar a comparação entre os modelos resultantes.

4.1 *Obtenção e preparação dos dados*

O conjunto inicial de dados utilizados para o processo de treino e teste dos classificadores foi obtido a partir de três diferentes trabalhos. O primeiro, formado por 141 questões, foi criado por Mohammed e Omar [6], e é constituído por questões abertas retiradas de *sites*, livros e trabalhos anteriores, e rotuladas de acordo com as seis classes da taxonomia de Bloom. A segunda base de dados, também disponibilizada por Mohammed e Omar [6], é composta por 600 questões, e foi obtida a partir de outro trabalho citado pelo autor.

Por último, o terceiro conjunto de dados foi proposto por Basu et al. [8], onde foi observado que uma classe da base de dados TREC[1] poderia ser mapeada para os quatro primeiros níveis da taxonomia de Bloom (Conhecimento, Compreensão, Aplicação e Análise). Assim, esta última base de dados é composta por 1284 questões.

Desta maneira, foram obtidas 2025 questões classificadas de acordo com a taxonomia de Bloom e em seguida foi efetuada uma mistura (*shuffle*) dos dados, para distribuir as questões aleatoriamente na base de dados. Após este procedimento foi realizada a separação dos dados, resultando em 80% das questões para serem utilizadas no treinamento dos modelos de classificação (1620 questões), e 20% para a base de dados de teste (405 questões). A distribuição das classes na base de dados completa pode ser observada na Figura 2.

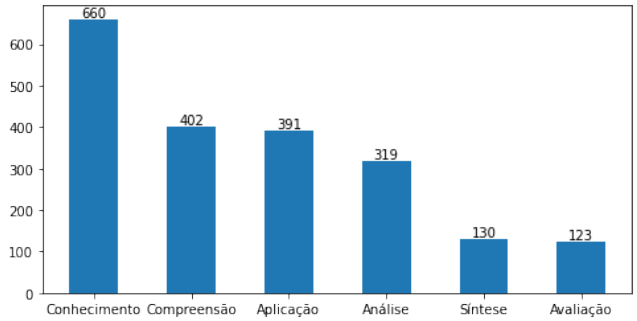

Figura 2: Distribuição de classes na base de dados.

4.2 *Pré-processamento dos dados*

Para efetuar a preparação dos dados, utilizados no treino e teste dos modelos de classificação, foram utilizadas estratégias para o processamento das questões. Inicialmente foi realizada a remoção de pontuações contidas nas sentenças, remoção de números, e remoção de *stop words*. Após a realização deste processo, foi efetuada a lematização das palavras contidas nas questões disponíveis na base de dados. O processo de lematização foi preferido em detrimento do de *Stemming* devido ao processo que o algoritmo utiliza para realizar a contração dos termos. O algoritmo de lematização leva em consideração o contexto no qual a palavra está inserida, sendo assim, é capaz de distinguir entre palavras que possuem significados diferentes. Já o algoritmo de *Stemming* efetua a contração dos termos sem analisar o contexto ao qual está inserido.

4.3 *Treino e teste dos classificadores*

Inicialmente, foi utilizado o *TF-IDF* para a extração de *features* dos documentos dispostos na base de dados. Desta forma, as questões foram convertidas de linguagem natural para medidas numéricas, que representam a relevância da palavra dentro da base de dados. Assim, o algoritmo de *TF-IDF* extraiu 3049 *features* para cada instância da base de dados de treino e de teste.

Em seguida, foi efetuado o treinamento dos modelos que utilizam estratégias de *Gradient Boosting* para melhorar a performance da classificação (*Extreme Gradient Boosting (XGBoost)* e *CatBoost*), na qual até o momento não foram encontrados trabalhos utilizando tais abordagens para a classificação de questões de acordo com a taxonomia de Bloom. Além disso, foram implementados outros modelos utilizando os algoritmos *SVM (Support Vector Machine)* e *Random*

[1]Base de dados TREC - https://rdrr.io/cran/textdata/man/dataset_trec.html

Forests, para serem utilizados como referenciais de comparação (*baselines*). Por fim, os modelos resultantes foram comparados.

4.4 *Aumento da base de dados*

Foi utilizada uma abordagem conhecida como *Data Augmentation* para aumentar a base de dados de treino dos classificadores. Assim, foi possível observar a influência deste procedimento nos modelos treinados, sem influenciar os resultados com a modificação das questões utilizadas para testes.

Antes da realização do processo de aumento de base de dados, foram executados todos os passos de pré-processamento descritos na Seção 4.2, sendo eles: remoção de números e símbolos, remoção de *stop words* e lematização. Desta maneira, é possível realizar a busca por sinônimos apenas para as palavras mais relevantes.

Para realizar o procedimento de aumento de base de dados, foi utilizado o *corpus* de palavras WordNet[2], disponível pela biblioteca NLTK[3]. Assim, para cada sentença contida na base de dados de treino, foram criadas no máximo seis frases adicionais, formadas a partir da substituição de uma das palavras por um sinônimo fornecido pelo *corpus WordNet*.

Além disso, foi observado que a busca por sinônimos originava palavras que não faziam parte da mesma classe gramatical da palavra original. Desta maneira, foi criado um filtro para selecionar apenas sinônimos que pertencessem à mesma classe gramatical da palavra original, utilizando a ferramenta de *Pos Tagging* disponibilizada pela biblioteca *NLTK*. Por fim, foi obtida uma nova base de dados de treino, contendo 13949 questões, com a distribuição de classes apresentada na Figura 3, e foi observado que a maioria das sentenças que constituem a nova base de dados não sofreram mudanças de significado devido a utilização da técnica proposta.

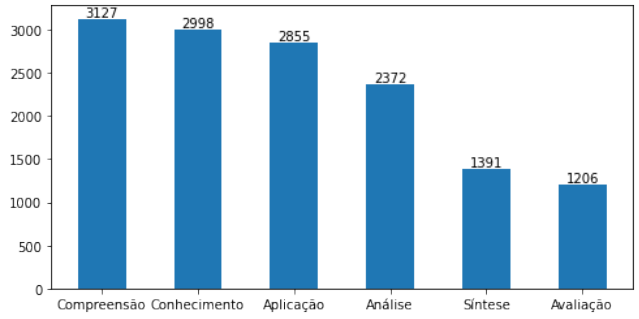

Figura 3: Distribuição de classes na base de dados aumentada.

Após este procedimento, os algoritmos utilizados na Seção 4.3 foram utilizados para a realização do processo de treinamento com a base de dados aumentada. Assim, as métricas de Precisão, Acurácia, *Recall* e *F1* foram obtidas para serem comparadas com o processo de treinamento anterior.

[2]WordNet - https://wordnet.princeton.edu
[3]Natural Language Toolkit (NLTK) - https://www.nltk.org

4.5 Métricas de avaliação para os modelos de classificação

Para efetuar a avaliação de modelos de classificação e a comparação entre os resultados obtidos, quatro métricas são frequentemente utilizadas, são elas: Precisão, Acurácia, *Recall* e *F1*.

- Precisão: A precisão é medida pela quantidade de VerdadeirosPositivos (i.e., classificações corretas da classe positiva) sobre a soma dos VerdadeirosPositivos e FalsosPositivos (i.e., erro em que o modelo previu a classe positiva, quando o real valor era negativa). Como descrito na Equação 2.

$$P = \frac{VP}{VP + FP} \tag{2}$$

- Acurácia: Indica uma visão geral dos resultados do modelo, e é definida como sendo a quantidade de VerdadeirosPositivos somada a quantidade de VerdadeirosNegativos (i.e., classificações corretas da classe negativa), sobre o total de questões, assim como mostrado na Equação 3.

$$A = \frac{VP + VN}{Total} \tag{3}$$

- *Recall*: É obtida através da quantidade de VerdadeirosPositivos dividida pela soma dos VerdadeirosPositivos e FalsosNegativos (i.e., erro em que o modelo previu a classe negativa, quando o real valor era positiva), ilustrada na Equação 4.

$$R = \frac{VP}{VP + FN} \tag{4}$$

- *F1*: É uma média entre a Precisão e o *Recall*. Calculada como sendo duas vezes a Precisão, vezes o *Recall*, dividido pela soma entre a Precisão e o *Recall* (Equação 5).

$$F1 = \frac{2 \times P \times R}{P + R} \tag{5}$$

5 RESULTADOS E DISCUSSÕES

Nesta seção, serão abordados os resultados e discussões acerca dos experimentos executados seguindo a metodologia descrita na Seção 4. Assim, os modelos de classificação treinados com os algoritmos *SVM* e *Random Forest* foram utilizados como *baseline* para a comparação com os modelos resultantes da utilização dos algoritmos *XGBoost* e *CatBoost*. Além disso, a técnica de aumento de base de dados foi utilizada como estratégia para o aumento da eficácia dos modelos.

5.1 Avaliação inicial dos modelos de classificação

O primeiro experimento foi realizado com a utilização dos algoritmos *SVM* e *Random Forest* como base, para serem comparados com o *XGBoost* e *CatBoost*. Sendo assim, os modelos de classificação foram treinados e as métricas Precisão, Acurácia, *Recall* e *F1* foram obtidas, e estão descritas na Tabela 1.

Tabela 1: Resultados iniciais dos classificadores

Modelos	Precisão	Acurácia	*Recall*	*F1*
SVM	79,57%	79,50%	79,50%	**79,15%**
Random Forest	72,76%	70,12%	70,12%	70,41%
XGBoost	76,00%	74,32%	74,32%	73,23%
CatBoost	77,32%	76,79%	76,79%	75,88%

Assim, foi possível constatar o desempenho superior do modelo treinado com o *SVM* em todas as métricas. Os modelos de classificação treinados com *XGBoost* e *CatBoost* possuem métricas de *F1* próximas do modelo *SVM*, sendo elas 73,23% e 75,88%, respectivamente. Diferentemente dos demais, o modelo *Random Forest* demonstrou um desempenho inferior, com a medida *F1* de 70,41%.

Desta maneira, é possível verificar que, como todos os modelos apresentam métricas de Precisão acima das demais, a grande maioria das classificações de questões propostas pelos modelos estão realmente corretas.

Além disso, é possível observar que a métrica Acurácia para todos os modelos permanece próxima às demais métricas. Indicando que a performance geral entre os modelos é semelhante.

5.2 Impacto do aumento da base de dados

Devido a pouca quantidade de dados disponíveis para o treinamento dos modelos de classificação, foi utilizada uma abordagem com técnicas para efetuar o aumento da base de dados de treino. Com a realização deste experimento foi produzida uma base de dados com 13949 questões a partir das etapas descritas na Seção 4.4.

Assim, o próximo experimento tem como objetivo validar o uso da técnica de *Data Augmentation* descrita na Seção 4.4. Desta maneira, foram utilizados os modelos treinados na Seção 5.1 para serem comparados com os modelos treinados com a nova base de dados. Para tal finalidade foram utilizadas as mesmas métricas da Seção 5.1 para a realização da comparação. Tais métricas, do treinamento com a base de dados aumentada, foram obtidas e estão apresentadas na Tabela 2.

Tabela 2: Resultados de treino com base de dados aumentada

Modelos	Precisão	Acurácia	*Recall*	*F1*
SVM	76,86%	76,79%	76,79%	76,31%
Random Forest	81,18%	80,74%	80,74%	**80,46%**
XGBoost	75,75%	75,06%	75,06%	74,03%
CatBoost	79,17%	78,51%	78,51%	77,71%

Sendo assim, é possível observar que todos os modelos apresentaram melhorias significativas com o aumento da base de dados, com exceção do *SVM* que apresentou F1 de 76,31% e Precisão de 76,86%, cerca de 3,40% pior que o resultado anterior. Não foi possível chegar a uma conclusão exata da causa desta perda de eficácia para o modelo *SVM*, assim, será estudada em trabalhos futuros com o intuito de entender melhor as classificações propostas pelo modelo.

Além disso, foi constatado que o algoritmo *Random Forest* apresentou um ganho de eficácia maior que os demais. Deste modo, a métrica de F1 resultante para o *Random Forest* foi de 80,46%, e a Precisão de 81,18%. Na Tabela 3, estão descritos os comparativos entre os resultados iniciais (sem o aumento da base de dados) e os resultados com o aumento da base de dados em termos de F1.

Tabela 3: Comparação entre métricas F1 dos modelos treinados com e sem aumento de base de dados

Modelos	s/ aumento de dados	c/ aumento de dados	Melhoria em %
SVM	79,15%	76,31%	-3,588%
Random Forest	70,41%	80,46%	14,27%
XGBoost	73,23%	74,03%	1,092%
CatBoost	75,88%	77,71%	2,411%

Por fim, utilizando a métrica F1 (média entre Precisão e *Recall*) podemos constatar que o modelo treinado com o algoritmo *Random Forest* sofreu um acréscimo de 14,27% no valor observado para a métrica F1. Desta maneira, esta medida representa um aumento na precisão das classificações e na quantidade de acertos para valores esperados como pertencentes a uma determinada classe.

Com isso, é perceptível que os algoritmos *SVM* e *Random Forest*, comparados com os demais, representaram uma escolha melhor para a classificação de questões de acordo com a taxonomia de Bloom, tendo em vista que todas as métricas do modelo treinado com o *SVM* sem o aumento da base de dados ainda possuem melhores resultados que as referentes ao *XGBoost* e *CatBoost* com o aumento dos dados. Porém, os algoritmos de *XGBoost* e *CatBoost* se mostraram concorrentes à altura dos demais, tendo em vista a melhoria das classificações refletida em todas as métricas com o aumento automático dos dados. Como exemplo desta melhoria podemos citar a métrica Precisão do modelo treinado com o *CatBoost*, que teve seus resultados melhorados em 2,392%, indicando que o modelo passou a acertar mais as predições para as classes da taxonomia de Bloom.

6 CONCLUSÕES E TRABALHOS FUTUROS

Neste trabalho, comparamos dois algoritmos muito utilizados para o processo de classificação de questões de acordo com a taxonomia de Bloom (*SVM* e *Random Forest*), com dois outros pouco utilizados para esta problemática, *XGBoost* e *CatBoost*. Além disso, foram abordadas questões de aumento de base de dados, que consideramos relevantes para o aumento da eficácia dos modelos de classificação.

Assim, foi constatado que a utilização de técnicas de *Data Augmentation* proporcionou um ganho de eficiência considerável, e, portanto, pode auxiliar os modelos de classificação a alcançarem melhores resultados. Sendo assim, acreditamos que o desenvolvimento deste trabalho tenha sido importante para motivar outros pesquisadores na escolha de abordagens para a rotulagem automática de questões. Além disso, tendo em vista que o treinamento de tais classificadores pode auxiliar o processo de rotulagem de bases de dados compostas por questões, professores podem se beneficiar com estas técnicas para entender melhor a dificuldade das questões que estão propondo para seus alunos.

Em trabalhos futuros, pretendemos obter uma base de dados maior para o treinamento e avaliação dos modelos. Para tal finalidade pode ser efetuada a anotação de base de dados já existentes, mas que ainda não são rotuladas, devido a escassez de bases classificadas com a taxonomia de Bloom. Adicionalmente, pretendemos utilizar outras abordagens para o aumento automático da base de dados utilizada para o treinamento dos modelos. Assim, serão consideradas inicialmente as seguintes técnicas: *Easy Data Augmentation*, proposta por Wei e Zou [9], e *Back Translation*, como foi utilizada por Xie et al. [10].

REFERÊNCIAS

[1] D. Abduljabbar and N. Omar. 2015. Exam questions classification based on Bloom's taxonomy cognitive level using classifiers combination. *Journal of theoretical and applied information technology* 78 (2015), 447–455.

[2] Clarissa Dirks Alison Crowe and Mary Pat Wenderoth. 2008. Biology in Bloom: Implementing Bloom's Taxonomy to Enhance Student Learning in Biology. *CBE—Life Sciences Education* 7, 4 (2008), 347–430. https://doi.org/10.1187/cbe.08-05-0024

[3] B. S. Bloom, M. B. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwohl. 1956. *Taxonomy of educational objectives. The classification of educational goals. Handbook 1: Cognitive domain.* Longmans Green, New York.

[4] Gobinda G. Chowdhury. 2003. Natural language processing. *Annual Review of Information Science and Technology* 37, 1 (2003), 51–89. https://doi.org/10.1002/aris.1440370103 arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/aris.1440370103

[5] Lion F. Gardiner. 1994. *Redesigning Higher Education: Producing Dramatic Gains in Student Learning*. Vol. 7. Graduate School of Education and Human Development. The George Washington University. 233 pages.

[6] Manal Mohammed and Nazlia Omar. 2020. Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec. *PLOS ONE* 15, 3 (03 2020), 1–21. https://doi.org/10.1371/journal.pone.0230442

[7] Dale H. Schunk. 2011. *Learning Theories: An Educational Perspective*. Vol. 6. Boston: Pearson, The University of North Carolina at Greensboro. 574 pages.

[8] SShyamal Kumar Das Mandal Syaamantak Das and Anupam Basu. 2020. Identification of Cognitive Learning Complexity of Assessment Questions Using Multi-class Text Classification. *Contemporary Educational Technology* 12, 2 (05 2020). https://doi.org/10.30935/cedtech/8341

[9] Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. arXiv:cs.CL/1901.11196

[10] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Unsupervised Data Augmentation for Consistency Training. arXiv:cs.LG/1904.12848

[11] Anwar Ali Yahya, Zakaria Toukal, and Addin Osman. 2012. Bloom's Taxonomy–Based Classification for Item Bank Questions Using Support Vector Machines. In *Modern Advances in Intelligent Systems and Tools*, Wei Ding, He Jiang, Moonis Ali, and Mingchu Li (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 135–140.