Marcos Alécio Spalenza

*p*Nota: Uma Análise das Estruturas Textuais para Avaliação de Respostas Discursivas Curtas

Vitória, ES

2021

Marcos Alécio Spalenza

*p*Nota: Uma Análise das Estruturas Textuais para Avaliação de Respostas Discursivas Curtas

Tese de Doutorado submetida ao Programa de Pós-Graduação em Informática da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do Grau de Doutor em Ciência da Computação.

Universidade Federal do Espírito Santo – UFES

Centro Tecnológico

Programa de Pós-Graduação em Informática

Orientador: Prof. Dr. Elias de Oliveira

Coorientador: Prof[a]. Dra. Claudine Badue

Vitória, ES

2021

Aos meus pais, Marcos e Sirlene,
por me ensinarem, desde cedo, o valor da educação.

Agradecimentos

Agradeço a Deus.

Agradeço aos meus pais, Marcos e Sirlene, por todo carinho e atenção.

Agradeço ao meu irmão, Murilo, por cada palavra de apoio e pela companhia incondicional.

Minha gratidão ao professor Elias pela disponibilidade, confiança no meu trabalho e pelo apreço dado a cada resultado obtido durante todo o período do mestrado e doutorado.

Agradeço à professora Claudine pelas importantes contribuições e por cada momento de incentivo, garantindo sempre caminhos e soluções aos problemas.

Agradeço também a todos os demais professores que tornaram esta tese possível, seja durante as aulas, por meio de sugestões ou colaborações em experimentos e artigos.

Agradeço a cada um dos amigos e amigas que fiz durante todo o período de pós-graduação. Sem dúvida a presença de cada um foi essencial para finalização deste trabalho. Em especial, gostaria de destacar todos os companheiros de laboratório por cada um dos momentos compartilhados.

Agradeço a todos os membros do LCAD e do PPGI, por ter me acolhido, por prontamente atender todas as demandas necessárias e por todos esses anos de aprendizado.

Agradeço a Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES) pelo incentivo a pesquisa e desenvolvimento científico (processo 80136451).

Agradeço a *NVIDIA Corporation* pela doação de uma *NVIDIA TITAN V* através do *NVIDIA Academic Hardware Grant Program*, importante ao desenvolvimento desta e de várias pesquisas do laboratório.

Além disso, gostaria de expressar minha gratidão a todos os profissionais que trabalharam dioturnamente contra a Covid-19, buscando reduzir o impacto e as consequências sobre a população.

"O homem não é nada além daquilo que a educação faz dele."
(Immanuel Kant)

Publicações

Foram desenvolvidas as seguintes publicações no período do Doutorado que, em maior ou menor grau, apresentam certa relação com o tema deste trabalho.

Publicação de trabalhos em anais de congressos:

SPALENZA, M. A.; LUSQUINO-FILHO, L. A. D.; LIMA, P. M. V.; FRANÇA, F. M. G.; OLIVEIRA E. *LCAD-UFES at FakeDeS 2021: Fake News Detection Using Named Entity Recognition and Part-of-Speech Sequences. In: Proceedings of the Iberian Languages Evaluation Forum.* Málaga, Spain: CEUR-WS, 2021. (IberLEF - SEPLN 2021, v. 37), p. 646-654.

SPALENZA, M. A.; OLIVEIRA E.; LUSQUINO-FILHO, L. A. D.; LIMA, P. M. V.; FRANÇA, F. M. G. *Using NER + ML to Automatically Detect Fake News. In: Proceedings of the 20th International Conference on Intelligent Systems Design and Applications.* Online Event: Springer International Publishing, 2020. (ISDA 2020, v. 20), p. 1176-1187.

SPALENZA, M. A.; PIROVANI, J. P. C.; OLIVEIRA, E. *Structures Discovering for Optimizing External Clustering Validation Metrics. In: Proceedings of the 19th International Conference on Intelligent Systems Design and Applications.* Auburn (WA), USA: Springer International Publishing, 2019. (ISDA 2019, v. 19), p. 150-161.

SPALENZA, M. A.; NOGUEIRA, M. A.; ANDRADE, L. B.; OLIVEIRA E. *Uma Ferramenta para Mineração de Dados Educacionais: Extração de Informação em Ambientes Virtuais de Aprendizagem In: Computer on the Beach.* Florianópolis (SC), Brasil: Universidade do Vale do Itajaí - UNIVALI, 2018. v. 9, p. 741-750.

Participação em trabalhos publicados em anais de congressos:

OLIVEIRA, E.; SPALENZA, M. A. ; PIROVANI, J. P. C. *rAVA: A Robot for Virtual Support of Learning.* In: Proceedings of the 20th International Conference on Intelligent Systems Design and Applications. Online Event: Springer International Publishing, 2020. (ISDA 2020, v. 20), p. 1238-1247.

SILVA, W.; SPALENZA, M. A.; BOURGUET, J. R.; OLIVEIRA, E. *Towards a Tailored Hybrid Recommendation-based System for Computerized Adaptive Testing through Clustering and IRT. In: Proceedings of the International Conference on Computer Supported Education.* Libon, Portugal: SCITEPRESS, 2020. (CSEDU 2020, v. 12), p. 260-268.

SILVA, W.; SPALENZA, M. A.; BOURGUET, J. R.; OLIVEIRA, E. *Recommendation Filtering à la carte for Intelligent Tutoring Systems. In: Proceedings of the International Workshop on Algorithmic Bias in Search and Recommendation.* Online Event: Springer International Publishing, 2020. (BIAS 2020, v. 1), p. 58-65.

PIROVANI, J. P. C.; ALVES, J. SPALENZA, M. A.; SILVA, W.; COLOMBO, C. S.; OLIVEIRA, E. *Adapting NER (CRF+LG) for Many Textual Genres. In: Proceedings of the Iberian Languages Evaluation Forum.* Bilbao, Spain: CEUR-WS, 2019. (IberLEF - SEPLN 2019, v. 35), p. 421-433.

PIROVANI, J. P. C.; SPALENZA, M. A. ; OLIVEIRA, E. *Geração Automática de Questões a Partir do Reconhecimento de Entidades Nomeadas em Textos Didáticos In: XXVIII Simpósio Brasileiro de Informática na Educação.* Recife (PE), Brasil: Sociedade Brasileira de Computação, 2017. (SBIE 2017, v. 28), p. 1147-1156.

Resumo

O processo de avaliação é uma etapa fundamental para a verificação de aprendizagem e manutenção do andamento do ensino conforme o currículo previsto. Dentro da avaliação de aprendizagem, as questões discursivas são comumente utilizadas para desenvolver o pensamento crítico e as habilidades de escrita. Porém, com mais alunos, torna-se necessário ao professor o desenvolvimento de seus métodos, sem tornar a avaliação um fator limitante. Alinhado a isso, ressaltamos a quantidade de material para avaliação, mesmo que individualmente as respostas representarem pequenas quantidades de texto produzido. Portanto, o professor precisa avaliar cautelosamente todos os alunos para identificar possíveis problemas no aprendizado. Além disso, a adesão de métodos de suporte educacional tende a melhorar a qualidade dos materiais e impactar diretamente no desenvolvimento do aluno. Deste modo, neste trabalho apresentamos um modelo avaliativo usando *Active Learning*, ou seja, combinando os resultados de *clusterização* e classificação para apoio ao tutor na avaliação de respostas discursivas curtas. Através do reconhecimento das estruturas textuais de forma gramatical, morfológica, semântica, sintática, estatística ou sequencial identificamos padrões textuais em cada conjunto de respostas por questão. Assim, com os modelos de resposta anotados, ajustamos o modelo avaliativo para se aproximar das expectativas de nota do professor. Deste modo, apresentamos a robustez do modelo avaliativo produzido pelo sistema através de diferentes *datasets* da literatura. Em 255 questões, com um total de 65875 respostas, alcançamos *Accuracy* média de 72,11% e *F1* ponderado de 70,63% em relação aos avaliadores humanos.

Palavras-chaves: Avaliação Automática de Questões Discursivas. Aprendizado Ativo. Sistemas de Apoio ao Tutor. Processamento de Linguagem Natural. Classificação de Texto.

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis malesuada laoreet leo at interdum. Nullam neque eros, dignissim sed ipsum sed, sagittis laoreet nisi. Duis a pulvinar nisl. Aenean varius nisl eu magna facilisis porttitor. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Ut mattis tortor nisi, facilisis molestie arcu hendrerit sed. Donec placerat velit at odio dignissim luctus. Suspendisse potenti. Integer tristique mattis arcu, ut venenatis nulla tempor non. Donec at tincidunt nulla. Cras ac dignissim neque. Morbi in odio nulla. Donec posuere sem finibus, auctor nisl eu, posuere nisl. Duis sit amet neque id massa vehicula commodo dapibus eu elit. Sed nec leo eu sem viverra aliquet. Nam at nunc nec massa rutrum aliquam sed ac ante.

Keywords: Automatic Short Answer Grader. Semi-Supervised Learning. Tutor Support Systems. Natural Language Processing. Text Categorization.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AL	*Active Learning* (Aprendizado Ativo)
ML	*Machine Learning* (Aprendizado de Máquina)
NLP	*Natural Language Processing* (Processamento de Linguagem Natural)
SAG	*Short Answer Grader* (Avaliação de Questões Discursivas Curtas)
EDM	*Educational Data Mining* (Mineração de Dados Educacionais)
PR	*Pattern Recognition* (Reconhecimento de Padrões)
IR	*Information Retrieval* (Recuperação de Informação)
CAA	*Computer-Assisted Assessment* (Avaliação Assistida por Computadores)

Sumário

1 Introdução

A avaliação é uma etapa fundamental para o ensino, inclusive para garantir a eficiência dos processos de ensino-aprendizagem. Através do método avaliativo que o professor mensura a assimilação do conteúdo ministrado. Portanto, é por meio das avaliações que o professor observa o desempenho da turma e seu progresso nos conteúdos. Nesse aspecto, é fundamental que o professor verifique com frequência o aprendizado dos estudantes no decorrer da disciplina. Assim, essa aplicação permite ao professor interagir com os alunos junto com os materiais pedagógicos para reformulação e aperfeiçoamento da dinâmica de ensino. Deste modo, é com o acompanhamento da disciplina e o apoio ao educando que as atividades estabelecem meios de reformular e controlar o processo de ensino-aprendizagem (BARREIRA; BOAVIDA; ARAÚJO, 2006).

Através das atividades, somada ao acompanhamento em sala, avaliamos a proficiência dos estudantes sobre determinado domínio. A proficiência envolve avaliar o raciocínio segundo a capacidade de resolver problemas, tomar decisões e realizar inferências sobre o assunto (CASIRAGHI; ALMEIDA, 2017). O papel da avaliação, portanto, é diagnosticar, apreciar e verificar o aprendizado dos alunos para que o professor atue no processo de formação de modo a consolidar seu método de ensino (OLIVEIRA; SANTOS, 2005). Deste modo, além do desenvolvimento e da sequência de métodos avaliativos, são imprescindíveis a verificação dos resultados e a correção de problemas nos métodos de ensino-aprendizagem.

O método avaliativo é o que torna possível o acompanamento e a solução dos problemas com o aprendizado dos alunos. Essa identificação de problemas e as ações para contorná-los tornam a estrutura curricular personalizada, alinhando a turma de acordo com os objetivos da disciplina (BIGGS, 1998). Portanto, é através das atividades que criamos o modelo para mensurar o conhecimento individual dos alunos. Para isso, a mediação tecnológica consolidou-se para aplicação das atividades em quantidade e qualidade. Deste modo, os Ambientes Virtuais de Aprendizagem (AVA) (MAQUINÉ, 2020) se tornaram modelos virtuais para suporte das aulas para turmas presenciais e a distância (RAES et al., 2020). Com a mediação tecnológica, apoiamos o professor na criação, avaliação, recomendação e visualização de dados educacionais impactando diretamente no acompanhamento do currículo do aluno (PAIVA et al., 2012). Deste modo, é com as ferramentas de apoio que o tutor verifica a aptidão dos estudantes, de forma individual ou coletiva, para melhorar a adaptação e a experiência da disciplina.

Nesse ponto, o acompanhamento, a formulação, a aplicação e os resultados dos métodos avaliativos em meio computacional são estudados pela área de *Computer-Assisted Asssessment* (CAA) (BOGARÍN; CEREZO; ROMERO, 2018), ou em tradução literal

Tabela 1 – Exemplo de respostas curtas com amostras da questão Q2 do *dataset Powergrading*.

Powergrading	
Q2	
What is one right or freedom from the First Amendment of the U.S. Constitution?	
#	Resposta
fbccf723b6ca	freedom of speech
10814c63d220	freedom of speech
0704a8d6f8d9	to bear arm
256b545c9f10	free excess of religion
a83446496fcb	freedom of speech
83bbefc5bbae	freedom of speech.
f28ffbdde6b9	to bear arms
bc65e0296be8	freedom of speech.
3e1216d9295e	freedom of speech
830444330cd9	life
815667698f42	right to pursue happiness.

Avaliação Assistida por Computadores. Na literatura de CAA, existe uma extensa pesquisa por métodos avaliativos e sua aplicação de forma digital (PÉREZ-MARÍN; PASCUAL-NIETO; RODRÍGUEZ, 2009). Em especial, destacamos as questões discursivas curtas, fundamentais para o desenvolvimento da escrita para aplicação em ampla escala neste formato. A avaliação de questões discursivas é dispendiosa, demandando análises da relação de cada resposta com seu alinhamento com o tema. Portanto, o estudo de formas computacionais para suporte aos métodos de avaliação de respostas discursivas, garantem maior capacidade de aplicação e correção para mensurar o aprendizado.

As questões discursivas incluem vários modelos avaliativos, das questões de preenchimento até longas redações. Neste trabalho, entretanto, buscamos dar suporte ao professor na avaliação de respostas discursivas curtas. Ainda assim, com *datasets* de diferentes características de resposta e avaliação, definimos como respostas curtas conjuntos textuais de até 3 sentenças, compostos por até 100 palavras. Para caracterizar as respostas discursivas curtas apresentamos na Tabela 1 a atividade de exemplo *Powergrading-A2*.

A Tabela 1 apresenta o enunciado da questão e uma pequena amostra dos identificadores e respostas enviadas pelos estudantes. Vamos utilizar essa questão Q2 do *dataset Powergrading* como modelo de questão discursiva curta. Nesta questão temos respostas sucintas para um enunciado que direciona para um ou poucos caminhos de resposta que podem ser desenvolvidos pelos estudantes. Nas respostas selecionadas vemos diferentes direções seguidas pelos estudantes. Porém, segundo a constituição americana:

Congress shall make no law respecting an establishment of religion, or prohibi-

ting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the government for a redress of grievances.

Deste modo, todos que responderam algo descrito na Primeira Emenda estão corretos. Entretanto, como o próprio exemplo indica, os *datasets* são característicos pela liberdade de escrita e diversidade de conteúdos. A liberdade de escrita é caracteristica em questões abertas, ou seja, aonde há produção textual por parte do estudante. Assim, a liberdade de escrita é expressa pelas diferentes formas que os estudantes se referem a um ou mais modelos de resposta. Devido a essa liberdade, devemos formar representações do conteúdo que demonstram a perspectiva de cada um dos participantes do método avaliativo.

Na aplicação de questões discursivas é comum que exista uma expectativa de convergência entre respostas, de forma a não criar respostas únicas e pessoais. Designamos então como *modelo de respostas*, a tendência de determinados conjuntos de resposta de estudantes apresentarem entre si certo índice de similaridade. Por outro lado, em uma perspectiva avaliativa do professor, existe o alinhamento entre conteúdo ministrado em sala e conteúdo esperado nas respostas. Encontramos este alinhamento nas *respostas candidatas*, ou seja, respostas produzidas pelo professor para caracterizar a expectativa de resposta de acordo com o enunciado da atividade. Por fim, na perspectiva do sistema, existem os padrões textuais. Com a análise do texto através de métodos de reconhecimento de padrões, identificamos características que definem grupos de conteúdo em meio a diversidade textual. Denominamos tais grupos como *padrões de resposta*, pois caracterizam grupos de resposta ao sistema para compreensão do conteúdo e do tema abordado.

Cada uma das perspectivas sobre o conjunto de respostas é importante para designar as formas com a qual cada um no processo avaliativo observa a diversidade textual. Portanto, na visão dos alunos, temos a convergência das respostas na formação de modelos. Na análise do professor, os modelos têm viés avaliativo, e esse viés é representado através de respostas candidatas ou expectativas de resposta. As respostas candidatas são instâncias textuais, ou seja, são respostas de referência tal qual as produzidas pelos estudantes. Por outro lado, as expectativas de resposta são segmentos do conteúdos que demonstram traços da avaliação, caracterizando o processo avaliativo. E, por fim, na ótica do sistema, a equivalência entre as estruturas textuais indicam padrões, formados por grupos de resposta com as mesmas características.

Sabendo da complexidade do processo avaliativo, caracterizado pela análise de todo conteúdo, atribuição de notas e revisão de resultados, enfatizamos diversidade textual. A diversidade textual inclui as diferentes formas de linguagem adotadas. Isso inclui todas as possibilidades de atingir a expectativa de resposta do professor de acordo com o método

Tabela 2 – Média de observados nas respostas

Dataset	Características	Palavras	Caracteres
UNT	140	20	107
ASAP	2932	43	236
Beetle	98	10	50
Open University	418	10	55
Findes	123	8	40
Powergrading	178	4	20
PTASAG	906	13	72
SEB	110	13	64
VestUFES	1391	92	536

avaliativo. Assim como o professor, a automação do processo de análise textual inclui compreender as tendências de resposta de acordo com o conteúdo abordado. Deste modo, as características linguísticas do conjunto textual deve ser analisadas tal qual o realizado por um especialista, nesse caso o professor. Com essa perspectiva, o objetivo geral dos sistemas de CAA é reduzir o esforço avaliativo e a análise de toda diversidade textual. Para isso, os sistemas devem criar modelos linguísticos robustos, de forma a ter alto nível de compreensão linguística e interpretação textual. Essa robustez é indicada dado o potencial de identificar respostas distintas com mesmo viés de resposta, aproximando-se da linguagem dos professores e alunos.

Por isso, o reconhecimento das estruturas que formam a linguagem escrita são fundamentais para a descoberta de modelos, padrões e características relevantes nas respostas. Cada resposta, como um documento textual, é composta por uma série de características. Cada característica é extraída de acordo com a construção da estrutura textual, seja ela no aspecto gramatical, morfológico, semântico, sintático, estatístico ou sequencial. Portanto, a identificação das características mais relevantes nas respostas que formam os padrões para o sistema. A Tabela 2 apresenta os valores médios da composição de cada *dataset* segundo suas características textuais.

Como a Tabela 2 destaca, mesmo sobre um tipo específico de questão, temos variações importantes no modelo de resposta. Apesar de todos os conjuntos se enquadrarem nas especificações de respostas com menos de 100 palavras ou 3 sentenças, são evidentes as diferenças entre *datasets*. Enquanto majoritariamente temos conjuntos com respostas bem concisas (menores que 20 palavras), temos VestUFES e ASAP mais descritivos e, possívelmente, com enunciados mais abstratos. Diante do escopo delimitado para as respostas, sobressaem alguns desafios na compreensão computacional da linguagem e dos métodos avaliativos. Deste modo, é fundamental a produção de modelos avaliativos computacionais que demonstrem fluência no método adotado. Portanto, modelos avaliativos designam sistemas que aplicam o método de forma similar ao professor, realizando a inferência avaliativa.

Por conta dos vários desafios nas técnicas de CAA, porpomos um modelo de *Active Learning* com análise das estruturas que compõe o texto em uma aplicação de reconhecimento de padrões. O sistema chamado *p*Nota, apresenta a combinação de técnicas de pré-processamento, clusterização, amostragem, classificação e produção de *feedbacks* para criar um modelo consistente em relação a expectativa do professor. Dentro das estruturas textuais, investigamos a construção de cada uma das sentenças sob diferentes perspectivas, avaliando aspectos gramaticais, morfológicos, semânticos, sintáticos, estatísticos e sequenciais que formam o conjunto de respostas. Sendo a atribuição de notas de interesse do professor, as respostas são comparadas diante de sua similaridade em diferentes modelos de classificação ou regressão. Deste modo, com o modelo proposto para avaliação de respostas discursivas curtas (ou *Short Answer Graders* (SAG), esperamos identificar padrões de resposta, reproduzir o método avaliativo e criar *feedbacks* descritivos para todos os participantes da disciplina (ARTER; CHAPPUIS, 2006; SPALENZA et al., 2016b).

1.1 Problema

Dentro da literatura dos sistemas SAG, encontramos determinados problemas que foram listados por autores durante os anos de evolução da pesquisa. Os problemas são amplos, onde se busca aprimorar gradativamente os modelos avaliativos para conseguir resultados cada vez mais adequados aos do professor (PADÓ; PADÓ, 2021). Assim, buscamos para além da redução do esforço de correção, a ampliação da análise e a proximidade entre modelos avaliativos. A expectativa é a redução de erros nos modelos avaliativos, com o modelo do avaliador automático aproximando-se do modelo do professor, enquanto especialista no tema. Nesse aspecto, mesmo que seja um trabalho realizado há decadas, a literatura dos modelos SAG descreve demandas importantes e pouco estudadas até o momento.

Nos primeiros sistemas, a modelagem de questões discursivas era um trabalho realizado com o texto bruto (PÉREZ-MARÍN; PASCUAL-NIETO; RODRÍGUEZ, 2009). A partir disso, a busca por equivalência entre a resposta esperada e o texto dos estudantes falhou por inúmeras vezes na padronização dos documentos e na identificação de sinônimos (LEFFA, 2003). O estudo dessa pesquisa fomentou inúmeras discussões em torno da identificação do conhecimento obtido pelas respostas escritas pelo aluno. Por conta disso, a robustez dessa análise é parte fundamental de boa parte dos algoritmos atuais em SAG (FILIGHERA; STEUER; RENSING, 2020).

Na principal revisão da literatura sobre os sistemas SAG (BURROWS; GUREVYCH; STEIN, 2015), os autores reúnem 37 trabalhos realizados na área. Durante esse estudo, o autor destaca o problema da profundidade do aprendizado, em tradução literal para "*depth of learning*", separando as atividades em dois grupos: de reconhecimento e de

recuperação. No Brasil, conhecemos discriminamos os dois tipos de questão como abertas e fechadas (GUNTHER; LOPES-JÚNIOR, 2012). Então, é característica dos modelos SAG a produção de modelos complexos de correção de questões de recuperação, interpretando computacionalmente o conteúdo de respostas curtas em textos de escrita livre.

Um problema diretamente associado ao modelo de interpretação textual, entretanto, é a busca de convergência entre as respostas. É uma dificuldade a extração do viés de resposta em questões discursivas factuais com múltiplos contextos. Além disso, existe ainda maior complexidade para lidar com questões que resultam em respostas opinativas, individuais ou subjetivas (BAILEY; MEURERS, 2008). Apesar do conteúdo, é esperado do sistema que, independente do conteúdo recebido do professor, lide com a liberdade de escrita do estudante e analize a convergência entre respostas na tentativa de recuperar padrões compatíveis (SAHA et al., 2018).

Em geral, para além do reconhecimento de padrões de resposta, ainda existe o alinhamento entre o conteúdo das respostas e o critério avaliativo. Esse fator se destaca pela referência utilizada na criação do modelo avaliativo (KRITHIKA; NARAYANAN, 2015). O professor, no papel de especialista, deve ser seguido segundo seu padrão avaliativo na tentando imitá-lo (JORDAN, 2012; FUNAYAMA et al., 2020). Para ilustrar esse aspecto podemos usar como exemplo a avaliação de plágios de forma negativa. Como é de interesse dos sistemas SAG seguir o modelo avaliativo do professor, mesmo que a essência do conteúdo seja coincidente com respostas próximas, o padrão específico reconhecido negativamente deve receber avaliação equivalente por conta do plágio segundo o método avaliativo do professor. Assim, apesar da escolha de palavras alinhadas com um determinado modelo de resposta, é essencial que o sistema forme vínculos entre padrões de avaliação e de respostas para criação de modelos avaliativos robustos (HIGGINS et al., 2014). Nesse aspecto destacamos que é fundamental para além de um sistema tradicional de reconhecimento de padrões a extração do critério avaliativo do professor.

Sabendo disso, o critério avaliativo deve atrelar componentes textuais ao método avaliativo. A aquisição desse modelo deve ser feita através da identificação da forma que o professor avalia uma série de respostas. Porém, uma série de trabalhos utiliza descritores do padrão avaliativo para representar a forma que o professor interpreta modelos de resposta dentre expressões regulares e regras até o quadro de *rubrics* (BUTCHER; JORDAN, 2010; MOHLER; BUNESCU; MIHALCEA, 2011; RAMACHANDRAN; FOLTZ, 2015). Porém, isso contrapõe a proposta de reduzir o esforço avaliativo do professor, se considerarmos a necessidade de produção de qualquer conteúdo extra sobre a avaliação (ZESCH; HEILMAN; CAHILL, 2015; HORBACH; PINKAL, 2018). A partir daí para remontar o critério avaliativo do professor, deve priorizar o uso de padrões de avaliação sem requisitar descritores ou chaves de resposta.

Dentro desse aspecto, durante a análise da compatibilidade entre os modelos

do sistema *FreeText Author* e do professor (BUTCHER; JORDAN, 2010), os autores elencaram seis problemas. O primeiro é a omissão dos padrões de avaliação. O segundo é a identificação da associação entre palavras e a sua conexão com o modelo avaliativo. O terceiro é a necessidade de identificação estrutural da sentença. O quarto é o tratamento de classificações incorretas, em especial por parte do especialista. O quinto é o conflito entre padrões corretos e incorretos. E, por fim, o sexto problema listado pelos autores é diretamente relacionado aos demais, indicando o problema de confiabilidade do sistema como avaliador e a interpretação inconsistente do conteúdo textual.

Na perspectiva da omissão dos padrões avaliativos, detalhamos a diferença entre o sistema ter o conhecimento dos padrões textuais e a avaliação de padrões desconhecidos. Padrões desconhecidos podem ter *outliers* que estritamente recebem um modelo próprio de avaliação (FILIGHERA; STEUER; RENSING, 2020). Entretanto, para além dos métodos aleatórios de amostragem, a avaliação de questões discursivas curtas *a priori* indica um problema de classificação desbalanceada (DZIKOVSKA; NIELSEN; BREW, 2012). Deste modo, destacamos a importância dos métodos de anotação direcionada a diversidade textual e uso de métodos de verificação da distribuição das amostras (MARVANIYA et al., 2018).

O segundo listado, em uma outra esfera avaliativa, estabelece a criação de modelos robustos entre termos e classes (RAMACHANDRAN; FOLTZ, 2015). Portanto, torna-se característico segundo os autores que os modelos SAG devem incorporar detalhes sutis da avaliação (HORBACH; PINKAL, 2018). Portanto, a relação termo-classe deve ser dinâmica e, apesar da avaliação ser passível de revisões e ajustes a qualquer momento, sempre extrair o modelo que melhor atenda às expectativas do professor (SPALENZA et al., 2016b).

Na sequência, o terceiro problema listado é de aspecto estrutural, observando cada resposta segundo os detalhes de sua construção. Por consequência, além da análise detalhada do modelo textual, é fundamental uma extensa capacidade analítica do conteúdo (SAHA et al., 2018). Portanto, além do nível textual desejamos que a análise seja feita em vários níveis, incluindo verificação morfológica, semântica e sintática de cada resposta (SAKAGUCHI; HEILMAN; MADNANI, 2015; RIORDAN; FLOR; PUGH, 2019; SAHU; BHOWMICK, 2020). Deste modo, incluimos neste aspecto, além de formas de maximizar a aquisição de informações em texto, a análise estrutural para compreensão da escrita das respostas. Somado a isso, algumas abordagens vão além e ainda exploram a conexão semântica entre respostas, questões e domínios (DZIKOVSKA et al., 2013; SAHA et al., 2019).

O quarto problema inclui o tratamento de classificações incorretas. É extremamente relevante aos sistemas SAG a construção de justificativas com fundamentos em referências textuais (FUNAYAMA et al., 2020). Assim, é recorrente a possibilidade de remontar

as componentes que levam a correção de cada resposta, sejam regras de associação de respostas, padrões de expressões regulares ou a extração de características textuais (CHAKRABORTY; ROY; CHOUDHURY, 2017; KUMAR et al., 2019). Para além da necessidade de justificativa, para cada nota atribuída, ainda ressaltamos a capacidade de identificar *outliers* (DING et al., 2020) e garantir que não se torne uma influência ao método avaliativo. Nessa linha, é importante que os modelos compreendam o conteúdo sem avaliações tendenciosas (AZAD et al., 2020), realizando uma análise ampla do conteúdo anotado.

De forma contínua ao quarto, o quinto problema compreende a identificação de incoerências nas avaliações. Entretanto, a incoerência é algo esperado desde que a divergência existe mesmo que entre dois humanos especialistas (ARTSTEIN; POESIO, 2008; PADÓ; PADÓ, 2021). Mas é essencial minorar a diferença cada vez mais entre o modelo do especialista e o modelo de avaliação automática (CONDOR, 2020). Nessa dinâmica, ressaltamos a importância em isolar comportamentos anômalos do método avaliativo para que não influencie no comportamento geral do modelo automático.

Por fim, a confiabilidade do sistema, tangenciando todos os demais citados, é tratada no último item. Superficialmente podemos associar este problema a divergência de notas entre avaliadores. Porém, em um aspecto amplo, a confiabilidade do sistema passa do reconhecimento do critério avaliativo à criação de justificativas de nota através de modelos descritivos de *feedback* (KUMAR et al., 2019). O papel dos modelos de *feedback* vai além de descrever o que o sistema observou na avaliação. Este declara a todos os participantes a relação entre as respostas, o reconhecimento do critério de avaliativo e cada anotação do professor (MARVANIYA et al., 2018). Portanto, a confiabilidade do sistema passa por todos os níveis, desde a aquisição de um critério de avaliação coerente até a representação do conhecimento.

Para além disso podemos ainda citar dificuldade em encontrar os *datasets* utilizados por trabalhos da literatura (BURROWS; GUREVYCH; STEIN, 2015). É muito comum encontrar trabalhos no qual os autores coletaram dados na própria universidade e não as tornam públicas. Além disso, em SAG uma base de dados adequada deve caracterizar o processo avaliativo do professor e constar com relevantes resultados na literatura. Assim, neste trabalho identificamos, testamos e descrevemos uma série de *datasets* na avaliação do método proposto.

1.2 Proposta

Neste trabalho apresentamos um método de avaliação de respostas discursivas curtas através da análise da estrutura textual para produzir modelos avaliativos complexos. Para seu desenvolvimento, identificamos os problemas mais comuns descritos na literatura

como deficiências dos sistemas SAG, apresentando uma proposta de solução. Cada um destes problemas é detalhadamente descrito na Seção 1.1. Portanto, a ideia é desenvolver uma estrutura de reconhecimento do critério avaliativo do professor estabelecendo a relação entre as respostas e as notas atribuídas.

Para atender as demandas encontradas nos trabalhos em SAG utilizamos de técnicas clássicas de *Educational Data Mining* (EDM) (ROMERO et al., 2010), *Machine Learning* (ML) (HAN; PEI; KAMBER, 2011) e *Natural Language Processing* (NLP) (JURAFSKY; MARTIN, 2009). Apesar do método ter fundamento em modelos linguísticos complexos e comportar questões em diversas linguagens, o avaliamos nas principais bases de dados em *inglês* e *português* da literatura. Dentre os *datasets* observamos 3 tipos de avaliações: notas ordinais, notas discretas e notas contínuas (MORETTIN; BUSSAB, 2010). Portanto, neste trabalho, estudamos estruturas para identificação das principais respostas do conjunto, reconhecimento do método avaliativo do professor (especialista) e elaboração *feedbacks*.

Para identificação das principais respostas apresentamos um modelo de aprendizado semi-supervisionado. No aprendizado semi-supervisionado o especialista ativamente passa o conhecimento para o algoritmo de classificação (BAEZA-YATES; RIBEIRO-NETO, 2011). O algoritmo, por sua vez, utiliza o as informações passadas para criar um modelo que imite o especialista na tarefa. Neste caso, o professor ensina ao sistema seu método avaliativo e, através da atribuição de notas, é formado um modelo que tenta replicar o método para as demais respostas da atividades (ROMERO et al., 2010). Cada uma das respostas enviadas para atividade é considerada uma amostra para o sistema. Dentre todas as amostras, é fundamental que o sistema aprenda cada uma das características das respostas, selecionando as principais por representatividade. Para essa seleção o sistema utiliza de técnicas de otimização e clusterização (EVERITT et al., 2011). As respostas selecionadas são denominadas de treinamento, pois serão utilizadas para produção dos modelos, enquanto as demais são o conjunto de teste.

No reconhecimento do método avaliativo do professor, modelos são criados para classificação das respostas discursivas. A categorização deve se aproximar ao máximo da tarefa realizada pelo professor, analisando detalhes parecidos na resposta. Portanto, o modelo avaliativo tem por premissa atender as expectativas do professor (PADÓ; PADÓ, 2021). Quanto menor a diferença entre a nota dada pelo sistema e a nota atribuída pelo professor, melhor o modelo criado. Consequentemente, os melhores modelos representam melhor a diversidade de notas e respostas com tendência menor de erros. Na gradação das notas, quanto maior a discrepância entre as notas mais críticos são os erros. Sabendo que, entre avaliadores humanos também existe esse erro (ARTSTEIN; POESIO, 2008). Os dados selecionados para treino do classificador ditam o conhecimento da gradação de notas distribuídas por ele. Portanto, o classificador recebe as características de cada resposta e a sua respectiva avaliação e as compara com as amostras de teste, com notas

não conhecidas. Portanto, o modelo de classificação, tomado aqui como avaliador, produz as notas complementares para o conjunto de dados de teste.

Por fim, a elaboração de *feedbacks* e relatórios é fundamental para o suporte ao professor. Em sala de aula, os *feedbacks* são um material que detalha a avaliação para professores e alunos e descrevem o método avaliativo de forma a sanar qualquer dúvida e evidenciar qualquer problema no aprendizado. Por outro lado, na perspectiva da interação do professor com o sistema, os *feedbacks* caracterizam a decisão, descrevem o modelo textual e a equivalência entre respostas. Portanto, em todos os ciclos do sistema esperamos reduzir o esforço de correção do tutor, apresentar resultados de alto nível com o modelo avaliativo e gerar materiais explicativos e completementares de qualidade.

1.3 Objetivos

O objetivo deste trabalho, portanto, é ajustar o modelo de correção criado pela máquina aos padrões estabelecidos pelo professor através da sua avaliação. Para isso, os modelos avaliativos devem compreender o método aplicado pelo professor, categorizando as respostas em classes, níveis ou intervalos contínuos de nota. Segundo a consistência de cada grupo, buscamos reduzir o esforço de correção do professor com a avaliação das respostas que apresentem apenas as principais caracteristicas textuais. Através de padrões bem definidos, esperamos reproduzir o critério avaliativo da questão justificando a classe atribuída através do seu respectivo sumário. Tal sumário, então, são os padrões de cada classe de nota partindo do agrupamento *a priori* das questões. É através desse sumário por nota que recuperamos um possível critério de correção. Desta forma, através do *p*Nota, esperamos que o professor esteja apto para gerenciar o seu método avaliativo em um tempo menor para concentrar-se na verificação de aprendizagem do aluno.

Portanto, temos como âmbito principal a criação de modelos para aproximar o critério avaliativo aplicado ao aluno da definição de padrões de correção e a criação de *feedbacks*. Para isso, estudamos os padrões avaliativos do professor e os métodos de representação do conhecimento encontrado em base de dados de questões discursivas curtas. Para atingir o objetivo geral descrevemos os seguintes objetivos específicos:

- Organizar os *datasets* públicos da literatura para estabelecer uma comparação com resultados obtidos em estudos correlatos (BURROWS; GUREVYCH; STEIN, 2015);

- Estudar o impacto das técnicas de Processamento de Linguagem Natural e Recuperação da Informação para a identificação da relação termo-classe de forma gramatical, morfológica, semântica, sintática, estatística ou sequencial (GALHARDI; BRANCHER, 2018; KUMAR et al., 2019; SAHU; BHOWMICK, 2020);

- Interpretar minuciosamente as respostas e o alinhamento do conteúdo, observando a frequência de ocorrência e co-ocorrência de termos segundo sua relevância (JORDAN, 2012; SAHA et al., 2018; DING et al., 2020);
- Elaborar e ajustar a avaliação de forma eficiente, assimilando o critério estabelecido pelo professor (ZESCH; HEILMAN; CAHILL, 2015; CONDOR, 2020; PADÓ; PADÓ, 2021);
- Criar modelos avaliativos robustos, associando as categorias de nota aos padrões textuais (BUTCHER; JORDAN, 2010; HEILMAN; MADNANI, 2015; BURROWS; GUREVYCH; STEIN, 2015);
- Identificar estruturas textuais para cada categoria de nota, removendo *outliers* e controlando da consistência da classificação (DING et al., 2020; FILIGHERA; STEUER; RENSING, 2020);
- Apresentar avaliações adequadas ao formato de correção do professor (HIGGINS et al., 2014; FUNAYAMA et al., 2020; PADÓ; PADÓ, 2021);
- Gerar *feedbacks* que colaborem com o processo avaliativo, como o quadro de *rubrics*, de forma a contribuir com a discussão de resultados e a representação do critério de correção (MARVANIYA et al., 2018; MIZUMOTO et al., 2019; SÜZEN et al., 2020).

1.4 Estrutura do Trabalho

A seguir são apresentados os conteúdos dessa tese. A proposta é discutida em detalhes através de 5 capítulos. Para além da Introdução, o trabalho é composto dos seguintes capítulos:

- **Capítulo 2 - Revisão de Literatura:** Apresenta uma breve revisão da literatura sobre métodos de análise e avaliação de respostas discursivas curtas.
- **Capítulo 3 - Método:** Define a estrutura do sistema *p*Nota e as formas utilizadas para efetuar de maneira abrangente a análise de respostas discursivas curtas.
- **Capítulo 4 - Experimentos e Resultados:** Descreve por meio de oito *datasets* as diferentes formas de apoio avaliativo, modelagem da relação termo-nota e a formação de *feedbacks* utilizados pelo sistema.
- **Capítulo 5 - Conclusão:** Discute as contribuições deste trabalho, conclusões extraídas dos resultados obtidos e as perspectivas de trabalhos futuros.

2 Revisão da Literatura

A sala de aula é um ambiente que produz diariamente grande quantidade de informações. As informações são essenciais para o acompanhar do aprendizado dos alunos, verificar a necessidade de reforço do conteúdo e monitorar o cumprimento do curricular. Tradicionalmente essa dinâmica faz parte dos métodos de ensino-aprendizagem empregados pelos professores, porém, superam a capacidade analítica dos mesmos (MADERO, 2019). Por conta disso, para ampliar a verificação do professor em analisar os materiais produzidos em sala, ganharam maior notoriedade e espaço prático os sistemas de EDM (SIEMENS; BAKER, 2012; ROMERO et al., 2010).

Em EDM, métodos de extração de informação são aplicados aos dados da classe de alunos para a aquisição de conhecimento, apoio ao tutor e acompanhamento do ensino (FERREIRA-MELLO et al., 2019). Através de técnicas de ML, ocorre a redução da carga do professor para tratamento e acompanhamento do conteúdo ministrado em sala. Deste modo, o professor torna-se responsável pela auditoria, monitoramento e aplicação dos resultados obtidos. Assim, os sistemas apoiam a descoberta de problemas de aprendizado, a personalização do ensino e a acompanhamento coletivo dos alunos em sala.

Portanto, através da mineração de dados, é possível ao professor a análise de todo material produzido pelos alunos, a criação de feedbacks individuais e a aplicação de reforço para determinados grupos de estudantes. Neste ponto, dentro dos métodos de EDM, um nicho de sistemas que tange diretamente essa demanda são os sistemas SAG (BURROWS; GUREVYCH; STEIN, 2015). Os SAG são responsáveis pela verificação em massa das respostas textuais curtas, auxiliando o professor no processo de correção. É característico deste tipo de questão a verificação do aprendizado do aluno segundo o material ministrado em sala (OLIVEIRA; CIARELLI; OLIVEIRA, 2013). Ao aluno, este tipo de questão é fundamental para prática da escrita, busca de informações e sumarização do conteúdo em poucas palavras. Portanto, este tipo de atividade envolve métodos relevantes para todos os níveis de ensino, principalmente durante o aprendizado e desenvolvimento da escrita (JOHNSTONE; ASHBAUGH; WARFIELD, 2002).

Apesar da relevância das questões discursivas curtas, sua aplicação é gradativamente reduzida pela alta carga-horária do professor em sala (BILGIN; ROWE; CLARK, 2017). Assim, torna-se uma demanda secundária o planejamento, a revisão e a análise do material dos alunos. O apoio computacional, reduz o tempo necessário fora da sala para avaliação do conteúdo, com o professor participando parcialmente do processo de avaliação (MING, 2005). O nicho dos métodos computacionais de apoio aos métodos avaliativos são conhecidos também por CAA (PÉREZ-MARÍN; PASCUAL-NIETO; RODRÍGUEZ, 2009). Neste

processo, os resultados obtidos são auditados pelo professor para garantir que o modelo avaliativo foi seguido fielmente para que a representação do conhecimento das respostas atenda coerentemente as demandas da atividade. Enquanto isso, a aplicação de técnicas de ML reflete que a descrição do modelo de correção utilizado é um potencial *feedback* com aplicação direta em sala (BUTCHER; JORDAN, 2010).

Para o uso dos métodos de SAG, é importante que a questão seja elaborada com o objetivo de analisar os conhecimentos dos estudantes segundo um domínio específico (BAILEY; MEURERS, 2008). Comumente separamos as questões em discursivas e objetivas, de acordo com o modelo de resposta esperada. As questões discursivas (BEZERRA, 2008) envolvem a liberdade de escrita do aluno, avaliando sua capacidade de descrição e desenvolvimento textual. Por outro lado as questões objetivas desenvolvem o raciocínio, a leitura e interpretação do material didático e a busca de informações. Sabendo disso, as questões discursivas permeiam ambos os tipos de habilidades do aluno. A Figura 1 caracteriza as atividades segundo os modelos de resposta (BURROWS; GUREVYCH; STEIN, 2015).

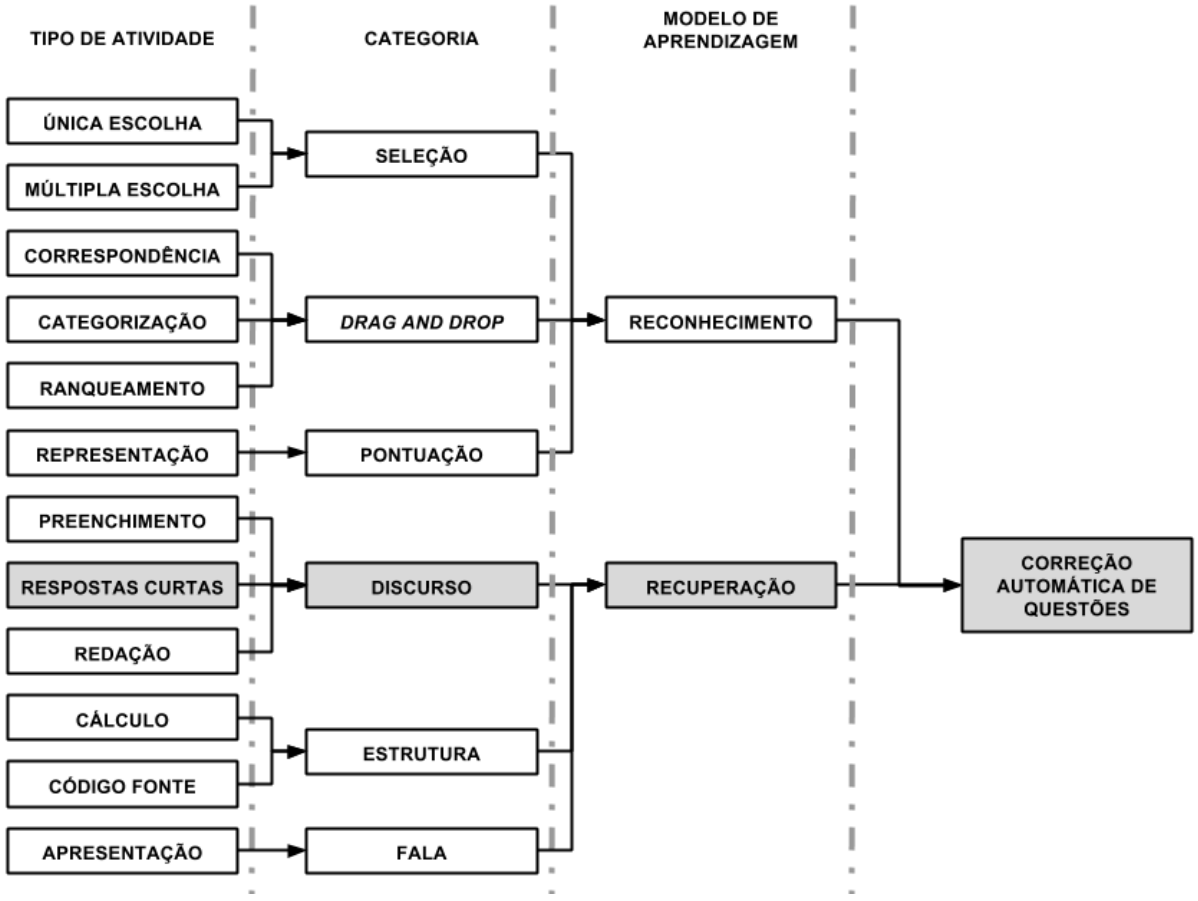


Figura 1 – A extração da informação e os tipos tradicionais de atividade aplicadas no cotidiano de sala de aula.

Como apresentado na Figura 1 o professor dispõe de alguns modelos de atividades que, refletem diferentes aspectos do aprendizado. Dentre as redações de cunho aberto e

irrestrito e as respostas diretas com opções elencadas no enunciado, as respostas discursivas encontram-se em âmbito intermediário (BAILEY; MEURERS, 2008). As respostas curtas buscam que o aluno estabeleça relação entre o aprendizado com material didático e a sua descrição textual. Assim, dentre os conhecimentos gerais, a questão deve evitar abordar temas de cunho interpretativo e que tangenciam experiências específicas de cada aluno (SIDDIQI; HARRISON, 2008). Por outro lado, a resposta deve representar a informação completa da questão, dando ao sistema embasamento para correção, evitando informações restritas ou codificadas (DING et al., 2020). A Figura 2 demonstra como o espectro de questões trabalhados através das respostas discursivas curtas.

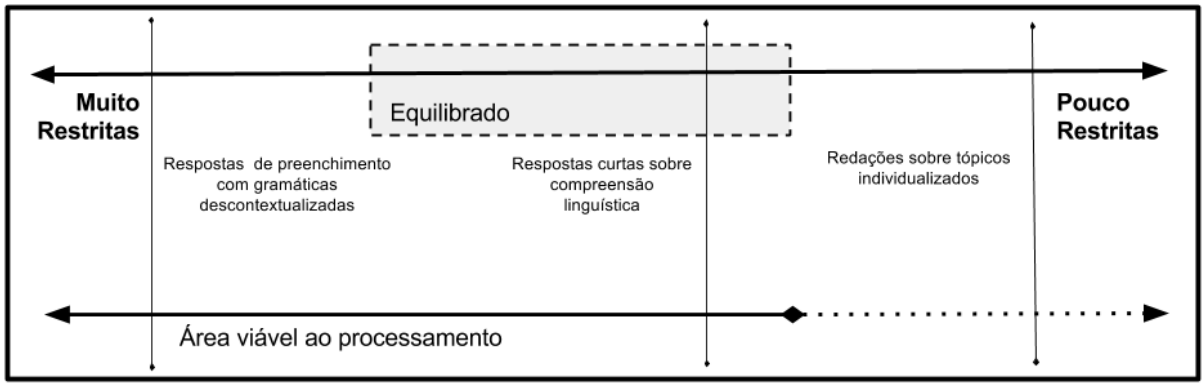


Figura 2 – Extração de informação em questões discursivas: entre respostas pequenas não-convergentes e a subjetividade das competências na avaliação de redações.

A Figura 2 caracteriza exatamente, dentro do nicho de questões discursivas, a dinâmica de uso do processamento computacional das respostas por parte do professor. O ideal é que a questão direcione o aluno a uma ou poucas respostas, evitando várias respostas corretas divergentes (SÜZEN et al., 2020) ou questões longas e subjetivas como redações (ALMEIDA-JÚNIOR; SPALENZA; OLIVEIRA, 2017). Quando as respostas não são únicas e apresentam um conhecimento comum não abstrato, é o ideal para uso dos métodos de correção automática. Portanto, é fundamental a convergência das respostas, para que as respostas apresentem uma ou poucas direções a serem abordadas pelos estudantes (FILIGHERA; STEUER; RENSING, 2020). Para isso, é fundamental que o sistema realize três passos. O primeiro é o aprendizado do modelo de respostas do aluno (RAMACHANDRAN; CHENG; FOLTZ, 2015). O segundo é através do modelo de respostas reconhecer o padrão avaliativo do professor (FUNAYAMA et al., 2020). Por fim, no terceiro passo, o sistema deve replicar o modelo avaliativo e elaborar *feedbacks* coerentes (FOWLER et al., 2021).

2.1 Avaliação Semi-Supervisionada

O método de aprendizado é o procedimento que dita a forma de aquisição de informações do sistema para criação de modelos com desempenho similar ao humano. Neste trabalho apresentamos um método de amostragem semi-supervisionado de aprendizado através da anotação do professor em itens selecionados através da *clusterização* (HORBACH; PINKAL, 2018). Porém, a requisição de anotação do professor para amostras de respostas não é o método tradicional para modelo avaliativo.

A grande maioria dos trabalhos utiliza amostragem através do particionamento entre treino e teste dos dados, previamente selecionado nos *datasets*. Considerando cada resposta dos estudantes uma amostra, o particionamento em treino e teste reflete a divisão *a priori* do conjunto de dados em um grupo para criação do modelo e outro para avaliação (HEILMAN; MADNANI, 2015). Esse modelo clássico permite ao sistema observar apenas uma parcela dos dados, onde o sistema realiza a inferência nos demais dados desconhecidos. Assim, o sistema deve absorver o modelo avaliativo do conjunto de treino e replicar o método avaliativo no conjunto de teste, pressupondo a equivalência dos mesmos. Porém, o modelo não necessariamente é similar ao de teste, não refletindo diretamente a aplicação de um sistema SAG em conjunto com o professor (SUNG et al., 2019).

Outros métodos, mais próximos da demanda do professor, utilizam de exemplos anotados de respostas para criação de modelo (BANJADE; RUS; NIRAULA, 2015; ROY et al., 2016). Tais exemplos são denominadas respostas candidatas. As respostas candidatas, são amostras elaboradas pelo professor e anotadas para representar seus padrões avaliativos. Os sistemas SAG com base nesse tipo de dado buscam, em geral, a comparação direta entre as respostas e o índice de sobreposição (KAR; CHATTERJEE; MANDAL, 2017; JIMENEZ; BECERRA; GELBUKH, 2013). Porém, este tipo de treinamento gera uma tendência na avaliação, com limitada interpretação das respostas dos alunos (RAMACHANDRAN; FOLTZ, 2015). O modelo criado não é capaz de identificar múltiplos contextos e as referências apresentadas pelo aluno. Portanto, as limitações da informação passada são um contraponto à liberdade textual esperada das atividades de escrita livre (BURROWS; GUREVYCH; STEIN, 2015). Além de tornar-se engessada, não necessariamente são similares aos demais documentos do *dataset*.

Para contornar as limitações, ainda existem alguns métodos utilizados para ampliar a capacidade de interpretação do sistema. O primeiro método visa maximizar o uso de informações das atividades. Nesta proposta, os sistemas são treinados com conteúdos adjacentes à questão, como o enunciado, o material de apoio e o quadro de *rubrics* utilizado pelo professor na correção (RAMACHANDRAN; CHENG; FOLTZ, 2015; WANG et al., 2019). O enunciado e o material de apoio adicionam ao sistema conhecimento externo sobre o tema. Enquanto as respostas candidatas e o quadro de *rubrics* são materiais descritivos do modelo avaliativo do professor para todos, inclusive o sistema (MIZUMOTO et al.,

2019; MARVANIYA et al., 2018). Por outro lado, existem sistemas que demandam modelos mais complexos do método avaliativo, como regras de avaliação e filtros de conteúdo feitos manualmente (PRIBADI et al., 2017; BUTCHER; JORDAN, 2010).

Outra estratégia é o uso de aumento de dados. Com aumento de dados as amostras passadas como treinamento são combinadas para representar de forma mais complexa o modelo avaliativo. O uso do aumento de dados torna os sistemas tradicionais um pouco mais robustos a alterações e mudanças nos padrões básicos, reduzindo a ocorrência de classificações tendenciosas (KUMAR et al., 2019; LUN et al., 2020). Assim, a quantidade de amostras para treinamento e variações para cada modelo de resposta torna-se muito superior à quantidade dada inicialmente. Outras formas incomuns ainda compreendem métodos de associação entre respostas com descoberta de padrões através de aprendizado não-supervisionado (ZHANG; SHAH; CHI, 2016). Neste conjunto de técnicas, destacam-se os métodos de *clusterização*. Com a *clusterização* os documentos de resposta são agrupados pelo coeficiente de similaridade e associados diretamente à uma determinada nota para o conjunto. Portanto, torna-se função do professor avaliar grupos de resposta segundo os componentes identificados como equivalentes (BASU; JACOBS; VANDERWENDE, 2013; ZESCH; HEILMAN; CAHILL, 2015).

De forma diferente das estratégias citadas, o aprendizado semi-supervisionado proposto combina os métodos de *clusterização* e classificação (OLIVEIRA et al., 2014). A *clusterização* é um conjunto de técnicas responsáveis por identificar de forma não-supervisionada um determinado número de agrupamentos de respostas pela similaridade. Os grupos, denominados *clusters*, indicam que os itens compartilham características equivalentes (EVERITT et al., 2011). Entretanto, na associação entre *clusterização* e classificação, os grupos são formados para amostragem, partindo dos *clusters* para reconhecimento da distribuição dos documentos. Essa amostragem visa identificar os itens que melhor descrevem cada agrupamento, associando as principais características textuais diretamente com o método avaliativo do professor.

2.2 Classificação de Documentos

Uma tradicional área em ML, a classificação de documentos, possui inúmeras subdivisões segundo a especialização, motivação e conteúdo do conjunto de documentos. As referências a cada conjunto de documentos podem ser dadas também como *dataset*, base de dados ou *corpus*. A coleção destes, porém, é denominada *corpora*. A classificação de documentos envolve treinar algoritmos de classificação com exemplos rotulados para replicar métodos de identificação de conteúdo e rotulação feitos por um especialista (BAEZA-YATES; RIBEIRO-NETO, 2011). Portanto, para além da origem e conteúdo dos documentos, o algoritmo deve se adaptar para especialização na triagem dos documentos

de acordo com suas características.

O especilista realiza uma leitura dos documentos e identifica informações específicas que justificam a categoria atribuída. Para replicar tal tarefa, através da análise do conteúdo, o sistema deve identificar características que estão diretamente relacionadas a cada classe de documentos. Dependendo da característica dos documentos, o conteúdo relevante de um documento para categorização pode incluir a identificação de poucas palavras-chave até a formação de modelos linguísticos complexos (JURAFSKY; MARTIN, 2009). Por exemplo, na triagem de documentos pré-formatados as informações básicas como título, autor e organizações ou setores responsáveis podem ser descritores diretos da classe a ser atribuída. Por outro lado, em modelos como SAG, é necessário que relações textuais complexas sejam avaliadas para atribuição de notas (PAIVA et al., 2012; YANG et al., 2021).

Deste modo, a atribuição de notas torna dos sistemas SAG uma complexa tarefa de classificação de documentos. É essencial a adaptação do algoritmo de acordo com o método de classificação utilizado pelo especialista. Portanto, apesar do conteúdo textual, a subjetividade do critério de avaliação deve ser levada em consideração pelo sistema (PADÓ; PADÓ, 2021). Assim, a combinação entre o reconhecimento do modelo avaliativo e o reconhecimento do modelo textual deve atender às expectativas do professor (CONDOR, 2020). Enquanto em parte das situações as notas fortemente correlacionadas com a ocorrência dos termos, em outras o critério do professor pode ter baixa correlação com os termos e apresentar diferentes nuances na atribuição de notas (AZAD et al., 2020). Deste modo, é determinante que o sistema compreenda a essência do conteúdo do documento enviado por cada aluno para reconhecimento da relação com as respectivas notas atribuídas (MOHLER; BUNESCU; MIHALCEA, 2011).

2.3 Processamento de Linguagem Natural

Para criação de um modelo linguístico, os sistemas utilizam estratégias de aquisição de informação com técnicas de NLP. As primeiras técnicas de SAG da literatura e os primeiros sistemas propostos utilizavam descritores (GALHARDI; BRANCHER, 2018). Os descritores são caracteríticas simples extraídas segundo o formato da escrita de cada documento. Em geral, são formados por características pré-definidas, de acordo com a estrutura da resposta do aluno, sem levar em consideração a profundidade do conteúdo (MOHLER; MIHALCEA, 2009). Dentre os descritores, os mais comuns eram a contagem de erros da linguagem, a quantidade de palavras e a frequência de certas classes gramaticais (RIORDAN; FLOR; PUGH, 2019; GALHARDI et al., 2018). Porém, as características pré-definidas, consequentemente, não atendem a uma grande quantidade de respostas, criando modelos linguísticos com pouca aderência ao conteúdo.

Posteriormente, observando os diferentes propósitos das questões discursivas curtas e sua aplicação multidiciplinar, surgiram estruturas para maior aquisição de informação e modelagem linguística (KUMAR et al., 2019; SAHA et al., 2018). Os modelos linguísticos ampliaram a aderência do sistema ao tema das atividades. Assim, através do conjunto de respostas, cada sistema elabora modelos linguísticos com contexto suficiente para encontrar associações entre palavras (TAN et al., 2020). Através dessas associações, os sistemas estabeleceram relações complexas entre os termos de cada resposta e o método de atribuição de nota do professor (SAHU; BHOWMICK, 2020).

As estratégias voltadas na análise do texto por completo, adicionaram muita informação aos sistemas. Porém, tais informações não necessáriamente são relevantes para o método avaliativo. Como consequência, ocorreu a evolução, desenvolvimento e uso de técnicas de ponderação, seleção de características e identificação de padrões textuais (BANJADE et al., 2016). Para ponderação textual o modelo mais comum é o Term Frequency - Inverse Document Frequency (TF-IDF) (BAEZA-YATES; RIBEIRO-NETO, 2011). O TF-IDF é um método clássico que realiza a ponderação de acordo com a frequência dos termos, equilibrando a relevância de cada termo segundo sua ocorrência nos documentos e no *dataset* (SULTAN; SALAZAR; SUMNER, 2016). Por outro lado, dentre as técnicas de seleção de características que se destacam, o *Latent Semantic Analysis* (LSA) (LANDAUER; FOLTZ; LAHAM, 1998) é uma das mais utilizadas na literatura (BASU; JACOBS; VANDERWENDE, 2013; SAHU; BHOWMICK, 2020). O uso desta técnica compreende identificar relações semânticas dentro do conjunto de respostas (MOHLER; MIHALCEA, 2009). Assim, através do LSA, os sistemas reunem o conteúdo que potencialmente contém maior significância no tema.

Entretanto, os modelos linguísticos criados através da frequência dos termos de cada resposta dos estudantes ainda não refletem uma análise complexa tal qual a do especialista. Portanto, na literatura existem estudos que propõe maior extração de informação textual, ainda que em textos curtos, para formação de componentes linguísticos mais robustos (SAHA et al., 2018; ZESCH; HORBACH, 2018). Uma estratégia é a análise estrutural dos termos, observando a construção frasal de cada resposta de aluno. Deste modo, na literatura alguns trabalhos citam a análise da construção gramatical das sentenças (RAMACHANDRAN; CHENG; FOLTZ, 2015; ROY et al., 2016).

Outras propostas porém, remontam o contéudo das respostas sob a perspectiva sequencial da construção textual (KUMAR; CHAKRABARTI; ROY, 2017). A análise, com a seleção de n termos de cada sentença da resposta é denominada *n-grams* (MANNING; SCHUTZE, 1999). Os sistemas avaliam as respostas através da vetorização das respostas com análise de compatibilidade entre essas sequências (SAKAGUCHI; HEILMAN; MADNANI, 2015; SULTAN; SALAZAR; SUMNER, 2016). Essas sequências subdividem cada resposta em pequenos trechos que contém de 1 a n termos para aplicar na análise de

equivalência e sobreposição entre respostas (JIMENEZ; BECERRA; GELBUKH, 2013).

Ainda nestes modelos, destacam-se propostas de seleção de características e filtragem de conteúdo (HIGGINS et al., 2014; SPALENZA et al., 2016b). Para filtragem de conteúdo, a identificação de termos comuns ou de baixa frequência representam um refinamento no modelo para análises mais consistentes do conteúdo (ZHANG; LIN; CHI, 2020; MARVANIYA et al., 2018). Termos comuns da linguagem em geral podem ser encontrados como *stopwords*, organizados em listas, são conectivos linguísticos muito utilizados que não têm aderência ao tema (JURAFSKY; MARTIN, 2009). Entretanto, em situação oposta, palavras com baixa frequência, com uso específico e, em geral, não são fundamentais para a resposta do aluno. Em ambos os casos, a filtragem propõe que termos de baixa correlação com o tema sejam removidos. Com uma proposta diferente, a seleção de características interpreta o conjunto de documentos em busca de termos correlatos. De acordo com a frequência de ocorrência e associação dentro do conjunto de respostas, termos são selcionados visando ampliar a capacidade do modelo avaliativo (KRITHIKA; NARAYANAN, 2015; SPALENZA et al., 2016a; HORBACH; PINKAL, 2018). Portanto, o intúito da seleção de características é diretamente relacionado ao modelo linguístico e avaliativo da base de conhecimento. Nessa perspectiva, apenas os termos selecionados são utilizados para representar o conjunto de respostas.

Recentemente, algo um pouco mais robusto do que a análise de vizinhança de termos vêm sendo empregada para avaliar a linguagem segmentos de resposta. Para isso, cada termo é avaliado por similariade no contexto ao qual é empregado. Um método em especial aplicado nesta proposta é denominado *word embeddings* (SUNG; DHAMECHA; MUKHI, 2019; GHAVIDEL; ZOUAQ; DESMARAIS, 2020). As *embeddings* são modelos linguísticos de grande dimensionalidade adquiridos de uma coleção de documentos (GOLDBERG; HIRST, 2017). Esses modelos relacionam o emprego de cada par de termos encontrados em coleções de larga escala. Assim, os sistemas avaliam a correspondência do emprego dos termos em cada sequência de forma pareada. Deste modo, os sistemas avaliam proximidade entre diferentes termos, frases e contextos de uso para cada resposta dos estudantes (RIORDAN et al., 2017).

2.4 Avaliação de Questões Discursivas Curtas

Os sistemas SAG para análise documental complexa são compostos por um conjunto de métodos que incluem a criação do modelo linguístico, organização do conhecimento e a identificação de características relevantes. Apesar disso, uma parte fundamental dos sistemas SAG são os classificadores de alta qualidade (FUNAYAMA et al., 2020). Portanto, são os classificadores que destacam o conhecimento adquirido nas etapas anteriores e o apredizado do modelo avaliativo (MOHLER; BUNESCU; MIHALCEA, 2011).

O propósito do classificador é compreender, replicar e descrever o modelo do professor (especialista) (YANG et al., 2021). Assim, é função do sistema identificar caracerísticas relevantes para assimilar a forma que o professor avalia cada resposta enviada pelos estudantes (JORDAN, 2012; MAO et al., 2018). Em geral, os avaliadores automáticos são divididos segundo quatro diferentes técnicas: por mapeamento de conceitos, extração de informação, análise de *corpus*, algoritmos de ML (BURROWS; GUREVYCH; STEIN, 2015).

O método de mapeamento de conceitos consiste em um processo de detecção de determinado conteúdo nas respostas produzidas pelos estudantes. O reconhecimento de conteúdo, portanto, é realizado com análise de alinhamento entre termos de respostas (JIMENEZ; BECERRA; GELBUKH, 2013). Deste modo, é fundamental neste método avaliativo, identificar a existência dos principais conceitos nas respostas para a atribuição de notas (KAR; CHATTERJEE; MANDAL, 2017; CHAKRABORTY; ROY; CHOUDHURY, 2017). Porém, mesmo com a construção automática de padrões através da amostragem, não é garantida a consistência dos modelos produzidos (AZAD et al., 2020). Deste modo, o principal fator destes sistemas é a busca por compatibilidade entre respostas, tornando o sistema muito dependente do objetivo da questão e o conteúdo enviado nas respostas (FILIGHERA; STEUER; RENSING, 2020).

Por outro lado, métodos de extração de informação apresentam características de identificação factual nas respostas dos estudantes. Portanto, compreendem métodos mais robustos de análise do conteúdo, sendo compostos por operações de reconhecimento de padrões e séries de expressões regulares (RAMACHANDRAN; CHENG; FOLTZ, 2015; BUTCHER; JORDAN, 2010). Assim, sistemas SAG com base na extração de informação apresentam modelos de resposta para análise da equivalência de cada resposta com a expectativa de resposta do professor. Deste modo, a associação entre respostas estabelece maior profundidade ao conhecimento do sistema sobre o conteúdo (TAN et al., 2020). Então, o modelo de avaliação utilizado pelo sistema torna-se próximo da observação do professor ao conjunto de respostas, porém, atendendo apenas modelos pré-definidos.

De forma distinta, os métodos baseados em *corpus* traçam análises estatísticas das respostas de cada conjunto de dados (KUMAR et al., 2019). Neste método, os sistemas utilizam de análises da linguagem para validação do alinhamento entre respostas, interpretar variações de uso e caracterizar o conteúdo das respostas (ZIAI; OTT; MEURERS, 2012; MENINI et al., 2019). Para além dos termos utilizados, a adição de informação acrescenta diversidade semântica, tornando modelos mais flexíveis para análise do vocabulário do material (FOWLER et al., 2021).

Apesar da consistência dos modelos anteriores, existem limitações em um âmbito geral da aplicação de cada uma das técnicas de acordo com um base de conhecimento (RIORDAN; FLOR; PUGH, 2019; DING et al., 2020). Em geral, as descrições de modelo

avaliativo do especialista não representam bem o conhecimento para a criação do modelo avaliativo do sistema (FILIGHERA; STEUER; RENSING, 2020). Em contraste aos modelos superficiais, as técnicas de ML foram incorporadas na análise textual para criação de modelos mais robustos, com fundamentação estatística (GALHARDI et al., 2018). Assim, modelos de aprendizado alinham o conteúdo dos documentos, através das diferentes componentes textuais, para reconhecimento dos padrões (SÜZEN et al., 2020). Portanto, os métodos criam estruturas mais complexas que regras, sendo capazes de avaliar formatos distintos de resposta (ZHANG; SHAH; CHI, 2016; SAHA et al., 2019; CAMUS; FILIGHERA, 2020). A robustez destes modelos permite a associação de padrões não convergentes, podendo estabelecer critérios distintos para amostras atribuídas a uma mesma nota.

Em geral, um objetivo dos sistemas SAG, descrito pela literatura, é mesclar os métodos e suas dinâmicas de aprendizado para evolução do modelo avaliativo (BURROWS; GUREVYCH; STEIN, 2015; ZESCH; HORBACH, 2018). Deste modo, é essencial a construção de modelos que comportem padrões avaliativos de alta qualidade e similares ao do especialista, reproduzindo com alta qualidade através de ML (JORDAN, 2012). Apesar das dificuldades e dos detalhes subjetivos da avaliação (ROY; RAJKUMAR; NARAHARI, 2018), o intúito é que o desenvolvimento do modelo avaliativo compreenda a relação entre diferentes características de avaliação e a capacidade de atender diferentes domínios (SUNG et al., 2019; SAHA et al., 2019). Portanto, espera-se o desenvolvimento de sistemas SAG mais robustos, lidando com diferentes combinações entre respostas e avaliações, aprendendo pela demanda do professor o domínio empregado.

3 Método

Neste trabalho, apresentamos um modelo de avaliação de respostas discursivas curtas através da análise da relação entre o conteúdo das respostas dos estudantes e o método avaliativo do professor. Acompanhando o desenvolvimento recente da literatura dos sistemas SAG (BURROWS; GUREVYCH; STEIN, 2015), identificamos pontos sensíveis e problemas descritos nestes estudos. Utilizamos como base os fundamentos de análise documental e modelagem do método avaliativo do tutor para a criação de uma proposta de sistema SAG. Através deste direcionamento, verificamos os principais métodos para análise das componentes textuais para elaborar um conjunto robusto de informações sobre cada resposta. Associamos ao conhecimento das respostas uma descrição do método de correção do professor. Com isso, esperamos construir modelos com o intúito de maximizar os resultados de acordo com o padrão de correções coletados junto ao professor.

Deste modo, apresentamos um sistema composto por quatro módulos. O primeiro módulo é o de coleta de dados, verificação textual, extração de informação e organização do conhecimento. Nesta primeira etapa, o sistema verifica cada resposta individualmente e aplica tratamentos textuais para padronização e extração de características. O resultado desta etapa é o conjunto de vetores de documentos padronizado para processamento. O segundo módulo é composto pelo particionamento de forma semi-supervisionada das amostras para reconhecimento dos padrões de resposta. Tal método analisa a representatividade de cada vetor de respostas, realiza a amostragem e coleta as notas do professor no papel de especialista na avaliação. A próxima etapa recebe os subconjuntos de respostas, uma parte com requisição da avaliação e outra separada para avaliação do próprio sistema. Com isso, o terceiro módulo compreende o reconhecimento do padrão de correções para as amostras e a reprodução do critério avaliativo. A reprodução do processo observado nas amostras selecionadas é dada através de técnicas de classificação e regressão, de acordo com o padrão de notas. Ao fim desta etapa, todas as respostas contém notas atribuídas, sejam dadas pelo professor ou pelo sistema de forma colaborativa. Por fim, com o conjunto de informações utilizadas durante os processos, a quarta etapa, produz históricos, relatórios e *feedbacks* para descrever com detalhes cada *dataset*.

Antes da execução do sistema, a criação de bases de dados compreende organização dos dados e o cumprimento dos padrões de leitura. Cada base de dados deve apresentar uma série de respostas discursivas curtas e um índice como referência a cada aluno. A origem destes dados podem ser arquivos estruturados ou Ambientes Virtuais de Aprendizagem (AVA). Os arquivos estruturados, são conjuntos de amostras de resposta delimitados de forma organizada em colunas para descrever uma comunicação entre sistema e professor, incluindo índice, resposta, nota e *feedback*. Por outro lado, os AVA são plataformas

utilizadas pelos professores para interação direta com o aluno. Podemos citar como exemplos de AVA o Moodle e o *Google Classroom*. O uso deste tipo de sistema ganha ainda mais notoriedade com o Ensina a Distância (EaD), entretato, não está restrito ao mesmo. Para isso, utilizamos de um *framework* de coleta, transferência e controle das atividades da sala virtual para processamento externo (SPALENZA et al., 2018). Portanto, é responsabilidade da aplicação a coleta as atividades no ambiente virtual, a transferência para um servidor de processamento e o envio de resultados para o professor. A Figura 3 apresenta o funcionamento do método de coleta de dados em diferentes plataformas de ensino.

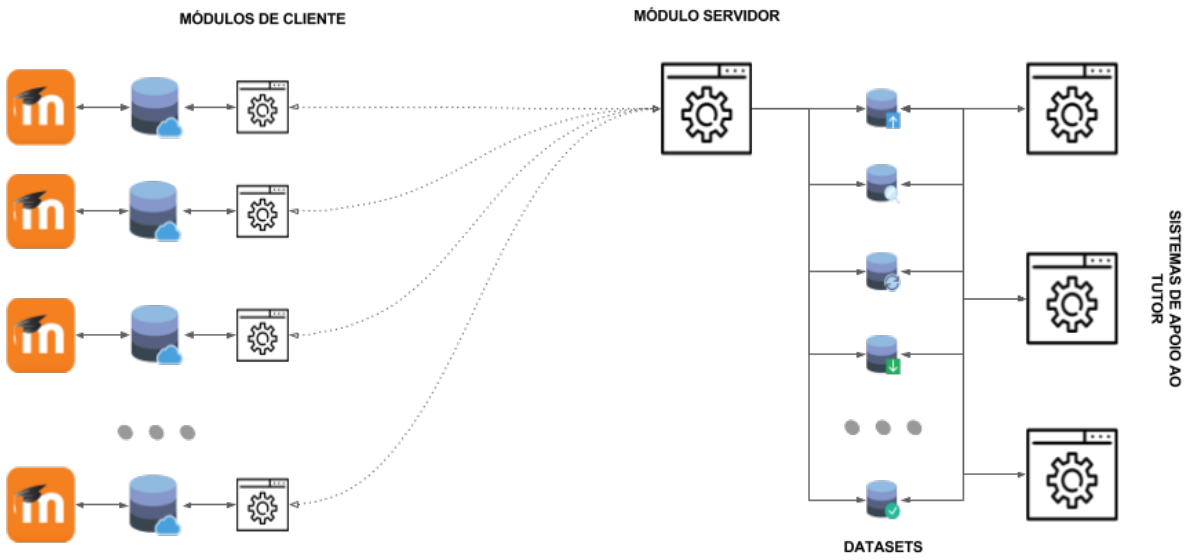


Figura 3 – *Framework* de transferência de dados, interligando plataformas AVA e as ferramentas de EDM.

A Figura 3 apresenta os métodos de extração de informação dos sistema com os AVA. Inicialmente, o módulo de transferência de dados é configurado em ambas as partes, no cliente AVA e no servidor de processamento. Com a configuração, o módulo acessa cada cliente e transfere as atividades ao qual os professores marcaram para análise. O *p*Nota avalia o conjunto de atividades e na primeira etapa realiza a requisição de anotação (avaliação) de amostras para treinamento do algoritmo de avaliação. O módulo de transferência envia marcações nas respostas requisitadas e o professor as avalia em seu ambiente de ensino. O sistema mantém sincronizada a versão da avaliação do professor e a versão no servidor. Em um segundo momento, com todas as respostas requisitadas já avaliadas, o sistema avaliador é treinado e recria o método de avaliação em modelos de classificação / regressão conforme as notas atribuídas. Os resultados, somando as notas atribuídas e os *feedbacks* gerados, são enviados para a plataforma de ensino novamente.

O professor, em qualquer momento fica aberto para finalizar/cancelar o processamento ao liberar a chave de buscas em sua plataforma. Da mesma forma, a nota atribuída

pelo professor é considerada a correta, sendo objetivo do sistema atender seu modelo avaliativo. Assim, é livre ao professor a alteração e controle de qualquer nota mesmo que ainda em processo de análise do sistema. Portanto, o professor a todo momento fica responsável por monitorar o processo avaliativo e ajustar os resultados propostos pelo sistema. A análise textual, seleção de amostras, modelos avaliativos e materiais explicativos aplicados pelo *p*Nota em cada atividade são apresentadas detalhadamente em quatro etapas do processo de correção automática.

3.1 Extração das Componentes Textuais

A primeira etapa, denominada de extração de componentes textuais, compreende a análise do conteúdo textual para a extração de conhecimento. Com os documentos organizados, o primeiro processo é a leitura do modelo de dados. Foram observados 3 diferentes modelos de dados: um único arquivo para todo o *dataset* da atividade, um arquivo de resposta textual para cada estudante em uma coleção para a atividade e, por fim, uma estrutura por aluno que contém os arquivos enviados para a com o conteúdo textual do aluno para a atividade. Com o carregamento dos arquivos, cada aluno é representado pelo seu identificador, seja ele uma referência da plataforma de origem ou sua respectiva ordem na estrutura do *dataset*.

Após a leitura do conjunto de dados, o sistema realiza uma série de processos de padronização, segmentação, filtragem, transformação e vetorização dos documentos. As etapas são sequenciais e encadeadas para análise do conteúdo em diferentes níveis. A padronização é composto pela coleta do conteúdo da resposta, remoção de conteúdos extras que permeiam o texto e a garantia da equivalência na ocorrência de cada termo. A segmentação visa a construção de vetores de resposta, termo-a-termo, identificando séries de *tokens* através de uma heurística cada palavra que a compõe. A filtragem, a partir dos vetores de resposta, seleciona as palavras com potencial relação com o conteúdo. A transformação compreende extrair as estruturas das componentes textuais e modificar cada palavra para um token representante. Por fim, com os documentos padronizados, ocorre a vetorização, analisando a frequência de *tokens* ou séries de *tokens* para representarem o conteúdo da submissão do aluno.

3.1.1 Padronização

Com a extração do texto do documento enviado pelo aluno, o conteúdo, neste primeiro momento, está em estado bruto. No estado bruto o documento precisa ser normalizado seguindo um padrão para os diferentes espaçamentos, acentuação e pontuação. Além disso, é fundamental remover conteúdos não interpretáveis incluindo caracteres não alfanuméricos e *tags* (marcações). Portanto, esta etapa é composta pelos seguintes

processos:

- Remover acentuação;
- Remover caracteres não-alfanuméricos;
- Remover pontuação;
- Remover espaços extras;
- Remover marcações.

Após cada um dos processos o conteúdo do aluno está normalizado para as próximas etapas. Os sinais gráficos auxiliam na identificação, pronúncia e leitura dos termos. Porém, computacionalmente, os sinais gráficos não é relevante para identificação de cada termo. O inverso ocorre com marcações de arquivos estruturados. As marcações, apesar da interpretação computacional, não fazem parte do conteúdo produzido pelo estudante. Portanto, ambos os casos não adicionam semanticamente ao conteúdo das respostas.

3.1.2 Segmentação

A partir dos documentos em formato padrão, com o texto normalizado, é possível partir para análise detalhada do conteúdo. Segmentações comuns podem particionar o conteúdo por palavras, por caracteres, por frases ou por parágrafos. Cada particionamento tem um aspecto específico alinhado com as tarefas realizadas na sequência. Neste caso, a análise detalhada do conteúdo depende do particionamento do que foi obtido em segmentos de palavras. Cada segmento é denominado *token*. Os *tokens*, neste caso, representam as palavras separadas por uma heurística, que delimita de cada segmento. Uma heurística simples é a *tokenização* por espaçamento, porém, como é um método simples é sujeito a muitas falhas. Apesar disso, métodos com melhor desempenho compreendem a aplicação da linguagem e consideram formas específicas de pontuação ou divisões do estilo textual.

A segmentação é o método que transforma o conteúdo em uma lista de palavras. A sequência de palavras permite que os próximos níveis trabalhem a perspectiva de cada *token* desta lista ou sua vizinhança. É muito comum que, durante o processo, o documento seja manipulado de diferentes formas, inclusive passando várias vezes pela transformação de texto em lista de *tokens* e vice-versa. Deste modo, os *tokens* permitem que cada palavra seja trabalhada de forma independente, sem impactar no conteúdo adjacente.

3.1.3 Filtragem

A filtragem de conteúdo é uma etapa muito importante desse processo. Uma dificuldade da filtragem é o balanceamento para alcançar níveis desejáveis de aquisição

de informação. Portanto, como esperado, a filtragem estabelece uma grande perda de informação e redução do conteúdo dos documentos. Entretanto, vale ressaltar que a perda de informação inerente ao processo caracteriza uma melhoria na consistência e na equivalência dos documentos. Como é uma limpeza conduzida pelo sistema, as características removidas representam detalhes com baixa correlação com a essência de cada resposta.

Em geral, nem todos os termos de uma sentença fazem parte do núcleo de interesse para análise das respostas. Algumas palavras independem do contexto ao qual são empregadas e não são aderentes ao tema. Esse é o caso das *stopwords*. As *stopwords* são palavras que são empregadas na linguagem como conectivos e não representam o conteúdo passado. Assim, são extremamente importantes para a leitura e interpretação do contexto, mas não adicionam informação quando empregadas. Assim, a lista de *stopwords* é um método que restringe a frase a palavras com maior potencial de relação com o contexto e o tema da resposta do estudante. Outros métodos de filtragem também incluem a remoção de palavras com poucos ou muitos caracteres e com tendência a serem muito específicos, quando não se enquadram como *stopwords*. Assim, podemos incluir como parte deste processo as seguintes etapas:

- Remover palavras pequenas (menores que 3 caracteres);
- Remover palavras grades (maiores que 15 caracteres);
- Remover *stopwords*;
- Remover números.

A remoção de partes do conteúdo devem cautelosamente observadas para não impactarem na capacidade de análise do conteúdo. Um bom exemplo é a extração dos números de uma resposta. Este método nem sempre é utilizado pelo sistema, dada sua influência no conteúdo. De modo prático, a aplicação deste método impacta diretamente na capacidade de análise de respostas compostas por números ou datas.

No entanto, podemos destacar a relevância da filtragem de conteúdo através de exemplos de aplicação. Por exemplo, uma situação ao qual encontramos uma fórmula em meio ao texto e, após as padronizações, se tornam comuns os caracteres soltos em meio ao texto. Neste caso, a função que elimina os *tokens* com menos de 3 caracteres remove tal conteúdo descontextualizado e direciona o sistema para atentar-se no contexto descrito pelo aluno. Por outro lado, a função que elimina com mais de 15 caracteres também tem papel fundamental. Podemos tomar como exemplo respostas que o estudante insere uma série de *links* como fontes do que foi descrito na resposta. Sob a perspectiva do sistema os *links* são grandes palavras únicas e não-interpretáveis. Então, a remoção de palavras com uma extensão incomum visa eliminar resquícios de conteúdo como *links* que não foram retirados nos níveis inciais de tratamento por conterem caracteres alfanuméricos.

Nesta etapa, portanto, os filtros de conteúdos são métodos de redução de ruído, responsáveis por discernir quais termos podem ser extraídos de cada item de resposta. O ruído em meio ao texto pode ser um grande problema para o desempenho do classificador em relação a interpretação do conhecimento. Identificando apenas a essência de cada resposta, esperamos que o sistema tenha maior capacidade de interpretativa da relação entre respostas.

3.1.4 Transformação

Em meio aos métodos de padronização, uma importante etapa é a transformação da série de *tokens*. As transformações envolvem métodos complexos de NLP, treinados para classificação de cada *token* segundo sua função na linguagem. Neste nível de trabalho do texto, o texto original é particionado, sendo observado em diferentes perpectivas. Os diferentes níveis analizados nesta etapa é apresentado à seguir:

- Análise gramatical: *Part-of-Speech Tags* (POS-Tags);
- Análise semântica:*Named Entity Recognition* (NER);
- Análise morfológica;
- Modificação: *Stemming*;
- Modificação: *Lemmatization*;
- Modificação: Tipografia.

Cada uma das atividades aplica uma diferente transformação no texto. Os primeiros três níveis, analíticos, acrescentam diversidade na informação de cada *token* do texto. O primeiro, usando *POS-Tags* realiza a análise gramatical de toda sequência. O método *POS-Tag* classifica cada palavra segundo sua função no âmbito gramatical, dentre verbos, adjetivos, pronomes, dentre 17 categorias (MARNEFFE et al., 2021). Em nível semântico, o NER é uma atividade de identificação de entidades nomeadas em meio ao texto livre (PIROVANI et al., 2019). Através do NER, categorias de nomes são definidas conforme a instância ao qual ele representa. Dentre as categorias reconhecidas neste trabalho estão *pessoa* (PER), *local* (LOC), *organização* (ORG) e *diversos* (MISC). Por último, o analisador morfológico identifica detalhes na construção de cada palavra. Pela análise morfológica certas palavras são identificadas segundo sua flexão. Dentre as flexões classificadas por cada termo estão as nominais (como gênero, número e definição) e verbais (pessoa, modo, tempo, voz). Adicionalmente, esse módulo também realiza algumas classificações léxicas de pronomes, adjetivos e advérbios (MARNEFFE et al., 2021).

Com análises linguísticas complexas, cada *token* é observado em diferentes perspe-

tivas. Adicionalmente, para lidar com o texto em si, aplicamos três tipos de modificadores. Os modificadores alteram o texto original para adicionar mais uma padronização. Aqui, no entanto, a padronização torna a linguagem mais próxima da compreensão do sistema do que da linguagem humana. Inicialmente, o processo de modificação de tipografia (*case*) torna todo o texto em letras maiúsculas ou minúsculas. Ao realizar essa mudança o sistema define palavras equivalentes para uma mesma forma. Do mesmo modo, os processos de *stemming* e *lemmatization* extraem das palavras flexionadas suas formas básicas (BAEZA-YATES; RIBEIRO-NETO, 2011). A forma simplificada extraída através do processo de *stemming* é a raiz da palavra. Enquanto isso, a simplificação resultante do processo de *lemmatization* é o *lemma* da palavra, ou seja, sua forma sem flexões. Em ambos os casos os *tokens* são modificados e todas as palavras de uma mesma base são dadas como coincidentes.

A resultante desses processos é uma forte análise das componentes textuais de cada documento de forma a compreender a construção do texto (SPALENZA et al., 2020). Através dessas verificações textuais, o texto recebe adição de diversos modelos que, em conjunto, caracterizam a construção termo-a-termo de cada frase. Os módulos analíticos, ampliam a informação de cada documento, tornando as nuances da escrita uma variável de interesse. Enquanto isso, os demais módulos visam aumentar a compatibilidade entre documentos para que escritas similares sejam trabalhadas de modo uniforme. Assim, o sistema como avaliador é responsável por compreender que o *corpus* foi trabalhado em diferentes perspectivas. Os termos são uma referência que buscam alinhamento com a resposta esperada na avaliação. Por outro lado, a estrutura frasal representa diferentes níveis linguísticos da escrita do estudante.

3.1.5 Vetorização

A vetorização, como última etapa do pré-processamento, é responsável por extrair o modelo numérico de cada documento, permitindo mensurar a diferença ou equivalência para os demais itens da coleção. Deste modo, os documentos são representados por vetores numéricos segundo seu padrão de características. Cada uma das características é analisada conforme sua frequência de ocorrência em cada documento do *dataset*. A representação vetorial numérica de cada documento pela frequência é denominada *Term Frequency* (TF). Sendo a coleção de documentos $D = d_0, d_1, d_2, \ldots d_i$ e as características (*features*) encontradas nos documentos $F = f_0, f_1, f_2, \ldots f_j$. Portanto, para cada documento d na coleção D, contamos a frequência de cada *feature* f_j do vocabulário F. Deste modo, a forma vetorial do documento de índice i é dada por d_i, sendo o vetor composto pela frequência n de cada *feature* no documento $n_{i,j}$. Então, podemos representar cada documento em D por sua forma vetorial $d_i = n_{i,0}, n_{i,1}, n_{i,2} \ldots n_{i,j}$, usando TF.

Dada as diferenças entre a frequência de cada termo em cada documento, é aplicada a pondereção para equilibrar a relação de frequência. A ponderação é denominada *Inverse*

Document Frequency (IDF). O *Term Frequency-Inverse Document Frequency* (TF-IDF) estabelece a relação de que termos que ocorrem em muitos documentos têm menor relevância (BAEZA-YATES; RIBEIRO-NETO, 2011). A ponderação ocorre conforme a Equação 3.1.

$$TF - IDF = d_{i,j} * \log \frac{n_D}{n_{d_j}} \tag{3.1}$$

Portanto, o IDF é uma ponderação na frequência de cada *feature* no vetor $d_{i,j}$, segundo o total de documentos n_{d_j} que contém f_j em relação ao total de documentos da coleção D. Essa ponderação reduz a diferença numérica entre uma característica encontrada em todos os documentos para as características que estão em grupos de documentos. Assim, o uso deste modelo potencialmente delimita melhor características relevantes em avaliações com mais gradações de notas. Então, a aplicação deste modelo está diretamente associada à capacidade de identificação de características com alta correlação a grupos específicos de nota.

No método de vetorização, durante a verificação de frequência de cada característica, existe a preocupação de manter a relação de vizinhança entre os termos e sua construção sequencial. Assim, para preservar o aspecto textual em sequências e identificar características adjacentes com alta correlação, utilizamos a análise por *n-grams*. Através dos *n-grams*, em vez de cada documento ser representado por um vetor simples da frequência de cada característica, essa frequência é calculada segundo uma sequências de n termos. Sendo aplicado valores n de 1 a 5-*grams*, utilizamos sequencias de 1 até 5 termos para analisar comportamento de cada documento em cada uma de suas perspectivas textuais. Portanto, as diferentes componentes textuais identificadas através de *n-grams* em busca de padrões mais complexos e associações de termos fortemente correlatos ao método avaliativo (SPALENZA et al., 2020).

3.2 Particionamento do Conjunto de Respostas

A partir dos vetores de documentos, o sistema *p*Nota torna-se capaz de comparar itens de resposta segundo suas componentes textuais. Com as características em formato numérico, começamos a interagir com o professor em busca da criação dos modelos que relacionem os documentos com as notas atribuídas. Entretanto, para criação destes modelos, o sistema precisa receber avaliação de alguns documentos para estabelecer a relação entre o que é o conteúdo de cada documento e a nota ao qual cada um recebe. Apesar de que muitos sistemas realizam uma amostragem aleatória, o *p*Nota realiza uma amostragem baseada na distribuição dos vetores. A análise da distribuição dos vetores e suas características é dada por meio de métodos de *clusterização*.

3.2.1 Clusterização

A *clusterização* é realizada com a otimização segundo o *elbow method*. Esse método é designado por testar sequência de valores de parâmetros para identificar a melhor combinação de *clusters* segundo uma métrica de qualidade. Em geral, a métrica de qualidade é diretamente relacionada com o propósito de uso dos *clusters*. O algoritmo de *clusterização* utilizado é o *Agglomerative Clustering* (SPALENZA; PIROVANI; OLIVEIRA, 2019), um método hierárquico de agrupamento por proximidade. O *Agglomerative* compreende formar *clusters* agrupando item a item até que um limiar de proximidade seja alcançado dado um k número de *clusters* (EVERITT et al., 2011).

Dentre as métricas estudadas estão o *Calinski-Harabasz Score* (CHS) (CALIŃSKI; J., 1974), *Davies-Bouldin Score* (DBS) (DAVIES; BOULDIN, 1979), *Silhouette Score* (SS) (ROUSSEEUW, 1987) e *Sum of Squared Errors* (SSE) (MAIMON; ROKACH, 2005). Essas métricas são denominadas índices de validação interna e avaliam os agrupamentos sem considerar a anotação de cada amostra, ou seja, de modo não-supervisionado. Cada índice é uma heurística utilizada para mensurar, sob diferentes perspectivas, a qualidade dos *clusters* gerados em relação a outros agrupamentos em um mesmo *dataset*. CHS mensura a razão entre a dispersão dos itens intra-*cluster* e a dispersão extra-*cluster*. DBS é o indice que estabelece a relação entre a média de similaridade entre as amostras do *cluster* para a média de similaridade entre-*clusters*. O SS é a média entre as distâncias das amostras pertencentes a um *cluster* em relação às amostras do *cluster* mais próximo. Por fim, SSE é uma métrica que avalia o erro de cada amostra que compõe um *cluster* em relação ao seu centróide. O centróide é o ponto médio dos itens que constituem cada *cluster*. Portanto, o centróide é uma instância representante da dispersão dos itens no *cluster*, porém é um ponto artificial e não necessáriamente uma amostra que o compõe.

Para a avaliação de respostas abertas, consideramos que o ideal são as análises que balanceam os itens de cada *cluster* em relação aos *cluters* adjacentes. Por padrão escolhemos a análise de *silhouette*, para identificar os resultados de *clusterização* com maior separabilidade entre os *clusters*. A separabilidade indica se os *clusters* formados são bem definidos, consistentes e sem sobreposição. Nesse índice, valores próximos a 1,0 representam agrupamentos consistentes, com distância para o *cluster* mais próximo. Valores negativos, aproximando-se de -1,0, indicam aleatoriedade na associação entre *clusters* e amostras, com confusão entre as os agrupamentos. Por outro lado, valores próximos a 0,0 indicam sobreposição entre *clusters*, com itens no limiar de pertencer diferentes grupos.

Em relação a verificação dos coeficientes de SS, a otimização com *elbow-method* identifica no intervalo de busca qual maximiza o resultado do índice. A otimização utiliza *Gaussian Process* para redução das possibilidades de busca. Esse método analisa cada teste pela distribuição dos valores da métrica de qualidade como uma *gaussiana*, buscando pontos de máxima da função. A resultante é dada pelo melhor valor encontrado (SPALENZA;

PIROVANI; OLIVEIRA, 2019). O atributo de controle é o k, número de *clusters*. O intervalo de k é definido por valores de 2 até $2 * \sqrt{n}$, sendo n o número de amostras do *dataset* (HAN; PEI; KAMBER, 2011). Simultaneamente, para cada combinação de k são realizados testes com vinte métricas de distância.

- braycurtis
- canberra
- chebyshev
- correlation
- cosine
- dice
- euclidean
- hamming
- haversine
- jaccard
- kulsinski
- mahalanobis
- manhattan
- matching
- minkowski
- rogerstanimoto
- russellrao
- sokalmichener
- sokalsneath
- yule

A resultante da otimização em clustering é escolhida como o teste que apresenta a melhor coeficiente de SS e com maior número de *clusters* formados. Enquanto o SS avalia a separabilidade dos agrupamentos, o maior número de *clusters* formados indica, na perspectiva do modelo avaliativo, uma possível coincidência entre conteúdo e notas. O agrupamento selecionado é utilizado para amostragem em um percentual do conjunto de respostas disponíveis. Para avaliar qualitativamente os resultados de *clusterização* utilizamos os índices *Adjusted Rand Index* (ARI), *Normalized Mutual Information* (NMI) e *Clustering Accuracy* (CA) (SPALENZA; PIROVANI; OLIVEIRA, 2019).

O ARI é um índice que compara a similaridade entre dois *clusters* de dados, levando em consideração o rótulo atribuído a cada amostra segundo seu agrupamento. Desta forma, o ARI estabelece uma relação entre pares de grupos formados e os rótulos reais de cada amostra, considerando possíveis inversões nos agrupamentos resultantes. Em geral, essa similaridade (*rand-index*) poderia ser calculada sem os rótulos, mas neste caso o índice é ajustado pela expectativa de ocorrência das classes. Portanto, o ajuste é dado com a associação entre cada classe d_y e sua representação segundo o *cluster* c_i. A Equação 3.2 apresenta o cálculo realizado por cada um deste índices de *clusterização*.

$$ARI = \frac{RI - Expected_{(RI)}}{MAX_{(RI)} - Expected_{(RI)}} \tag{3.2}$$

$$RI = \frac{\sum_{i=1}^{C} d_y == c_i}{D}$$

De modo distrinto, NMI estabelece uma escala de correlação entre os rótulos das amostra e a associação de cada uma a um cluster. Nesse aspecto, NMI relaciona a categoria aos *clusters* formados segundo sua coesão, segundo a Mutual Information (MI) e a Entropia (H) obtidas entre *clusters* (c) e classes (y). Em destaque, H é dado pelo somatório das probabilidade de cada classe y no conjunto, enquanto MI é dada pela diferença entre a

entropia global $H_{(y)}$ e a entropia do grupo em análise $H_{(y,c)}$. Essa relação de NMI para MI e H é dada pela Equação 3.3.

$$NMI_{(y,c)} = \frac{2 * MI_{(y,c)}}{H_{(y)} + H_{(c)}} \tag{3.3}$$

Sendo:

$$MI_{(y,c)} = H(y) - H_{(y,c)}$$

$$H_{(y)} = \sum_{i=1}^{n} P(y_i) \log_2(P(y_i))$$

E, por fim, utilizamos a CA para avaliar a categorização por voto majoritário dos grupos formados. Assim, com CA verificamos se os *clusters* são formados majoritáriamente por uma classe e, verificar a separação entre as classes. A Equação 3.4 apresenta a contagem das amostras d com determinada classe y para cada um dos C *clusters* formados.

$$CA = \frac{\sum_{i=1}^{C} MAX(d_{(y,c_i)})}{D} \tag{3.4}$$

Com os índices de qualidade de *clusterização*, estabelecemos uma perspectiva dos resultados esperados com os grupos formados e as categorias atribuídas pelo professor. Em especial, os índices foram utilizados como parâmetro para mensurar o impacto de *clusterizações* homogêneas e aprimorar a amostragem para posterior classificação. Deste modo, podemos estudar formas de seleção mais robusta de amostras após a formação de *clusters*.

3.2.2 Seleção de Amostras

Com a formação dos *clusters*, buscamos identificar as principais respostas de cada agrupamento para coleta do modelo avaliativo do professor. A amostragem é realizada com a coleta de um percentual dos itens que compõe o *dataset*. Essa coleta analisa padrões de documentos de cada grupo, a fim de compreender como é dada a avaliação do especialista para cada um dos diferentes itens. As amostras são selecionadas conforme critérios específicos, descrevendo um padrão específico do *cluster*. Os sete critérios de seleção de amostras utilizados estão listados abaixo:

- Par de amostras de menor similaridade no *cluster*;
- Par de amostras de maior similaridade no *cluster*;
- Amostra com mais características do *cluster*;

- Amostra com menos características do *cluster*;
- Amostra com menor índice de *silhouette* do *dataset*;
- Amostra aleatória do maior *cluster*;
- Amostra aleatória do *dataset*.

As sete instâncias, seguem uma ordem de prioridade conforme o percentual de itens coletados. O par de maior e o par de menor similaridade compreendem os itens mais convergentes e mais divergentes que compõe cada *cluster*. Em uma diferente perspectiva, a coleta de amostras dos itens de maior e menor número de características indicam as respostas que foram mais extensas e mais concisas, respectivamente. A coleta destes itens indica a consistência do padrão reconhecido na atribuição dos *clusters*. Em *clusters* com apenas um par de itens, por exemplo, todos os quatro modelos de amostragem são dados para as duas únicas amostras. Portanto, após essa seleção, não necessariamente foram identificadas 6 instâncias por *cluster*. A composição de até o percentual de amostragem por *dataset* é dado de três formas, aplicada conforme a demanda do sistema. A forma padrão, e mais robusta, é dada através da análise de dispersão de cada item.

A análise de dispersão calcula o coeficiente de *silhouette* da amostra. Tal qual o SS, esse índice determina a razão entre a distância da amostra para os demais itens do grupo em relação aos itens do *cluster* mais próximo. Desta forma, esse método incrementa as amostras por dispersão até alcançar o percentual de amostragem selecionado. Uma outra opção de seleção é a escolha de amostras por balanceamento do tamanho dos *clusters*. Tal método determina que um item seja aleatoriamente selecionado, sendo a seleção ponderada de acordo com a quantidade de itens que compõe cada grupo. Por fim, caso seja descartada a análise para amostragem dos demais itens, a seleção é realizada aleatoriamente, coletando itens sem uso dos agrupamentos.

Terminando este procedimento de seleção de amostras, as representantes de cada grupo são enviadas para avaliação do professor no papel de especialista. O especialista é responsável por atribuir notas de acordo com seu método avaliativo. Fica a cargo do próprio sistema identificar como os padrões de nota estão alinhados com os padrões textuais das respostas. Para isso, as notas coletadas devem ser estudadas pelo sistema para identificar o alinhamento do modelo avaliativo com o conteúdo das respostas.

3.3 Modelo Avaliativo

O desenvolvimento do modelo SAG acontece depois que o professor analisa todos as requisições de anotação, avaliando-as. Para a criação de um modelo SAG, é fundamental o desenvolvimento de um padrão de associação entre termos e notas. Entretanto, identificar

os detalhes observados pelo professor na avaliação não é trivial. Através do conjunto de respostas, o *p*Nota compara respostas que receberam a mesma avaliação para identificar padrões correspondentes. Os padrões encontrados em uma mesma nota, indicam detalhes que provavelmente foram levados em conta pelo professor na hora da avaliação. Deste modo, o sistema visa recuperar a expectativa de resposta por nota diante do alinhamento entre as componentes textuais.

3.3.1 Classificação

O processo de classificação enquadra-se com notas ordinais e discretas. Em tais situações, os classificadores atuam relacionando os padrões de cada categoria de nota ao conteúdo textual das respostas não anotados pelo professor. Deste modo, de cada técnica age mensurando a similaridade entre os modelos conhecidos para os demais que o *dataset* contém. Para isso utilizamos cinco algoritmos: *K-Nearest Neighboors*, *Decision Tree*, *Support Vector Machine*, *Gradient Boosting*, *Random Forest* e *WiSARD*.

O *K-Nearest Neighboors* (KNN) é o algoritmo de classificação pela análise da vizinhança amostral. Através do KNN cada amostra é categorizada pela distribuição local dos seus k vizinhos. A atribuição do rótulo é por voto majoritário, atribuindo o mesmo valor a amostra não anotada. Diferentemente deste, o algoritmo *Decision Tree* (DTR) estabelece a equivalência entre amostras, sob uma perspectiva das características que as compõe. O DTR define grupos anotados com a mesma classe pelos limiar das características que a compõe, gerando regras de decisão. As regras, elaboradas automaticamente, delimitam as principais *features* segundo os valores de tendência de classe. Portanto, o processo de classificação acontece com a comparação de cada um dos itens dentre a cadeia de decisões na estrutura de árvore.

Outro tradicional algoritmo, *Support Vector Machine* (SVM), estabelece uma forma distinta de observar os dados. Os dados, em grupos por categoria, formam um *kernel*. O *kernel*, diferente do DTR, cria modelo espacial que delimita a diferença entre categorias. Então, cada amostra, é identificada segundo sua posição em relação ao limiar de características dado o modelo representante da classe. De forma similar é aplicado o algoritmo *Wilkes, Stonham and Aleksander Recognition Device* (WSD) (ALEKSANDER; THOMAS; BOWDEN, 1984; LIMA-FILHO et al., 2020), conhecido como *WiSARD* [1]. O algoritmo produz um modelo binário de características através do registro de padrões de características. Cada padrão é reconhecido em análise sequencial de um intervalo de bits pré-definido. O modelo binário criado é comparado com as respostas não avaliadas, categorizando-as pela similaridade entre padrões. Especificamente para este algoritmo, a conversão dos vetores TF-IDF em seu formato binários foi dada com 1 *bit* por característica, de acordo com a esparsidade observada em dados textuais (MANNING; SCHUTZE, 1999).

[1] wisardpkg - https://github.com/IAZero/wisardpkg

Assim, dado como pré-requisito de sua aplicação, o padrão submetido é dado pela existência (valor 1) ou não (valor 0) de cada característica na resposta.

Adicionalmente, dois modelos de *ensemble* foram aplicados. Os *ensembles* são técnicas que combinam vários classificadores mais simples para determinar áreas de decisão mais robustas. Os classificadores simples são denominados *weak learners*, em busca de detalhes na avaliação entre termos e classes. Nesse aspecto, *Random Forest* (RDF), é um algoritmo que combina o método tradicional *Decision Tree* com *subsets* de amostras. Deste modo, o RDF combina análises parciais do conjunto de dados para definir regras de decisão mais complexas sobre a distribuição de amostras. De forma similar, o *Gradient Boosting* (GBC) combina uma série de *Regression Trees* para otimização diferencial da função de perda (*loss*). Nessa linha, o GBC observa o gradiente da função de perda com *Logistic Regression*. Nesse aspecto, com uma série de amostragens, a técnica procura minorar o erro de classificação obtido com a calibração do modelo segundo uma sequência de *subsets*.

A combinação com modelos tradicionais e técnicas de *ensemble* visa potencializar a capacidade analítica do método. Com diferentes formatos de dados, a proposta deste trabalho testa diferentes modelos procurando o que melhor se adequa ao padrão avaliativo do professor. Nesse aspecto, o método de classificação é escolhido de acordo com a similaridade entre o modelo automático com o critério do professor (PADÓ; PADÓ, 2021). Para avaliar este aspecto utilizamos o coeficiente *kappa* quadrático (COHEN, 1960). As amostras são separadas em dois grupos acordo com o *Stratified K-Fold*, para mensurar a capacidade de cada algoritmo na categorização das amostras. As amostras são separadas conhecendo a distribuição entre as categorias e, a cada ciclo, cada avaliador é comparado segundo a paridade de seus resultados com o avaliador humano (ARTSTEIN; POESIO, 2008).

Na sequência, a qualidade de cada um é avaliada com 4 métricas. A *Accuracy* (ACC), ou acurácia, mensura a equivalência percentual entre as avaliações. A *Precision* (PRE), ou precisão, estabelece a razão entre a atribuição correta de rótulos e a quantidade de atribuições incoerentes da mesma categoria. De forma similar, a *Recall* (REC), ou revocação, estabelece a razão entre a atribuição correta de rótulos e os itens de determinada categoria que foram classificados de forma incorreta. Por fim, F1 é o balanceamento entre PRE e REC, observando simultanemente os erros de e para cada classe. Este último é o único que além da avaliação média da atribuição de cada nota (macro) também é observada conforme a representatividade de cada categoria ao número total de amostras (ponderado). A Equação 3.5 apresenta a fórmula de cada uma das métricas citadas para avaliação qualitativa dos algoritmos de classificação testados.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.5}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Na Equação 3.5, vemos as fórmulas para mensurar a qualidade dos classificadores. Nelas T refere-se aos casos verdadeiros e F aos falsos. Da mesma forma, P refere-se aos casos positivos e N aos negativos. Com isso, todos os classificadores são avaliados segundo sua qualidade para divulgação dos resultados de modo a apreciar o resultado obtido. Porém, como descrito no processo de seleção, apenas o resultado que apresenta maior adequação ao modelo do professor é dado como resultado. Comparando estatisticamente os modelos, a expectativa é enviar apenas os resultados que sejam mais adequados, reduzindo as falhas no critério avaliativo do sistema.

3.3.2 Regressão

Em notas atribuídas em intervalos contínuos, são aplicados métodos de regressão. A regressão estima valores segundo o intervalo conhecido através das amostras. Os métodos buscam compreender o ajuste dos dados segundo a distribuição, elaborando o modelo minimizando o erro preditivo. Os cinco métodos de regressão aplicados são a *Regressão Linear*, *Lasso*, *K-Nearest Neighboors*, *Decision Tree* e *WiSARD*.

A Regressão Linear (LNREG) é um algoritmo que avalia a tendência linear das amostras segundo sua distribuição. Essa tendência linear busca, no espaço n-dimensional das características, definir os coeficientes de reta que minimizam o resíduo entre as amostras. É importante para o algoritmo determinar uma função de tendência dos dados. Minimizar o erro através dos coeficientes da função reflete na simplificação do conjunto de dados. Entretanto, é determinante que o modelo não apresente *overfitting* e um baixo desempenho com o viés dos dados de treinamento. Por outro lado, como espera-se do algoritmo, a aquisição de informação deve extrair um modelo que minimamente descreva os dados conhecidos, evitando a ocorrência de *underfitting*. Assim, o modelo simplificado deve ser direcionado ao desempenho linear e não apenas à associação forte com o conjunto de treinamento. Também é utilizada uma variante do LNREG tradicional, denominada *Least Absolute Shrinkage and Selection Operator - Lasso* (LSREG), que utiliza a normalização dos dados com a função $L1$, reduzindo a complexidade do modelo de dados e previnindo o *overfitting*.

Os demais três modelos, são similares aos modelos utilizados na classificação. O

K-Nearest Neighboors (KNREG), assim como o algoritmo de classificação, observa a distribuição dos dados e define o valor resultante de acordo com a vizinhança. Assim, o resultado de cada amostra de valor desconhecido é a interpolação entre os valores das K amostras mais próximas conhecidas. De forma semelhante, *Decision Tree* (DTREG) observa características semelhantes entre amostras e, por equivalência, divide em subgrupos. A subdivisão dos itens na árvore e o particionamento em subgrupos delimita regiões específicas com resultantes correspondentes por aproximação. Desta forma, após o particionamento das regiões amostrais em zonas de decisão, o valor dado para todas as amostras ali categorizadas é a média conhecida do subgrupo de treinamento. De forma similar funciona a WiSARD (WSREG), organizando registradores com as notas das respostas similares atribuindo o valor médio do registrador para respostas de padrão equivalente.

Segundo o modelo de notas contínuo, o método de avaliação dos métodos de regressão são dados através do erro da predição em relação a nota esperada. Assim, para mensurar a diferença entre a expectativa do professor e a nota resultante do sistema utilizamos o *Mean Absolute Error* (MAE), o *Mean Squared Error* (MSE) e o *Root Mean Squared Error* (RMSE). O MAE, erro médio absoluto, mensura a resíduo absoluto entre a nota predita e a nota dada pelo professor. Em outras palavras, o MAE avalia as diferênças em módulo entre os valores obtidos, segundo o alinhamento de cada predição com a expectativa do professor. Enquanto isso, MSE ou erro médio quadrático, é uma medida do resíduo entre os valores com penalização dos erros absolutos. Assim, através do MSE erros maiores têm maior impacto no sistema quando comparados com erros de menor grau. Por fim, o RMSE ou raiz do erro médio quadrático, é a raiz quadrada do valor obtido no MSE, normalizando o erro obtido nesta métrica em relação à avaliação do professor. A Equação 3.6 apresenta a fórmula de cada uma das métricas utilizadas para avaliação dos métodos de regressão citados.

$$MAE = \sum_{i=0}^{D} |y_i - p| \tag{3.6}$$

$$MSE = \sum_{i=0}^{D} (y_i - p)^2$$

$$RMSE = \sqrt{\sum_{i=0}^{D} (y_i - p)^2}$$

Na Equação 3.6 apresentamos as fórmulas de avaliar o erro do modelo criado conforme a expectativa de nota. Assim, em cada fórmula dos itens para cada amostra i na coleção comparamos as notas atribuídas pelo avaliador humano y e pelo sistema p. Apesar de serem comuns os erros entre modelos computacionais e a expectativa do especialista, é crucial para um bom avaliador automático a proximidade entre os modelos.

Assim, é esperado ao sistema que o erro seja minimizado e, como descrito, seja capaz de lidar com diferentes situações. Portanto, a capacidade avaliativa do sistema e seu nível de interpretação da linguagem podem ser mais relevantes a longo prazo do que o erro apresentado em situações incomuns (outliers). Destacamos ainda que, devido o nível de subjetividade inato ao processo avaliativo, os erros são comuns em qualquer correção, inclusive entre dois especialistas humanos. Nos sistemas SAG, foram observados durante a correção entre professores até 0,66 pontos de divergência em notas de 0 a 5 ponto (MOHLER; BUNESCU; MIHALCEA, 2011). Em uma escala de 0 a 10 pontos, representaria 1,32 pontos de divergência entre avaliadores humanos.

Para seleção do regressor mais adequado utilizamos a correlação de *Pearson*. Independente do nível de erro o método selecionado deve cumprir gradações similares ao método de atribuição de notas do professor. Considerando todos os algoritmos aplicados como capazes de realizar uma avaliação, mesmo que de forma básica, o uso da correlação nos permite verificar se a gradação é equivalente ao que foi atribuído pelo professor no treinamento. Para mensurar isso antes da avaliação, particionamos as amostras anotadas com KFold. O algoritmo é treinado com dois terços das amostras coletadas e validado em um terço. Após a predição no *subset* de avaliação, o modelo com maior índice de correlação observado é dado como padrão. Os resultados são gerados para todos os regressores, mas é encaminhado ao professor como resultado o de melhor correlação conhecida.

3.4 Relatórios

Após a classificação, com a atribuição de notas de todos os alunos, os relatórios são ferramentas importantes para aplicação direta em sala de aula. Os relatórios visam descrever para professores, estudantes e até desenvolvedores como é efetuada a análise das respostas, os métodos de reconhecimento de padrões e a coerência entre avaliação e modelos de resposta. Sabendo que cada modelo de resposta é até o momento uma associação entre termos e notas identificada pelo sistema, é determinante no processo de relatórios a apresentação deste modelo como descritor do método avaliativo.

Para ilustrar a produção de *feedbacks* utilizamos a atividade *Sandstone* da *Open University* (JORDAN, 2012). Esta atividade inglesa foi extraída do curso de Introdução à Ciências [2]. A Figura 4 apresenta o enunciado da questão apresentada aos alunos.

A atividade apresentada na Figura 4 faz parte da coleção disponível na 4. Nela os alunos devem determinar como é formado o *arenito*, rocha sedmentar resultante da compactação gradual da areia. Neste conjunto o professor avaliou manualmente cada uma das 1798 respostas de forma binária, sendo 1 atribuído para resposta *correta* e 0 para

[2] Sandstone - Open University. Disponível em https://www.open.edu/openlearn/science-maths-technology/science/geology/sandstone

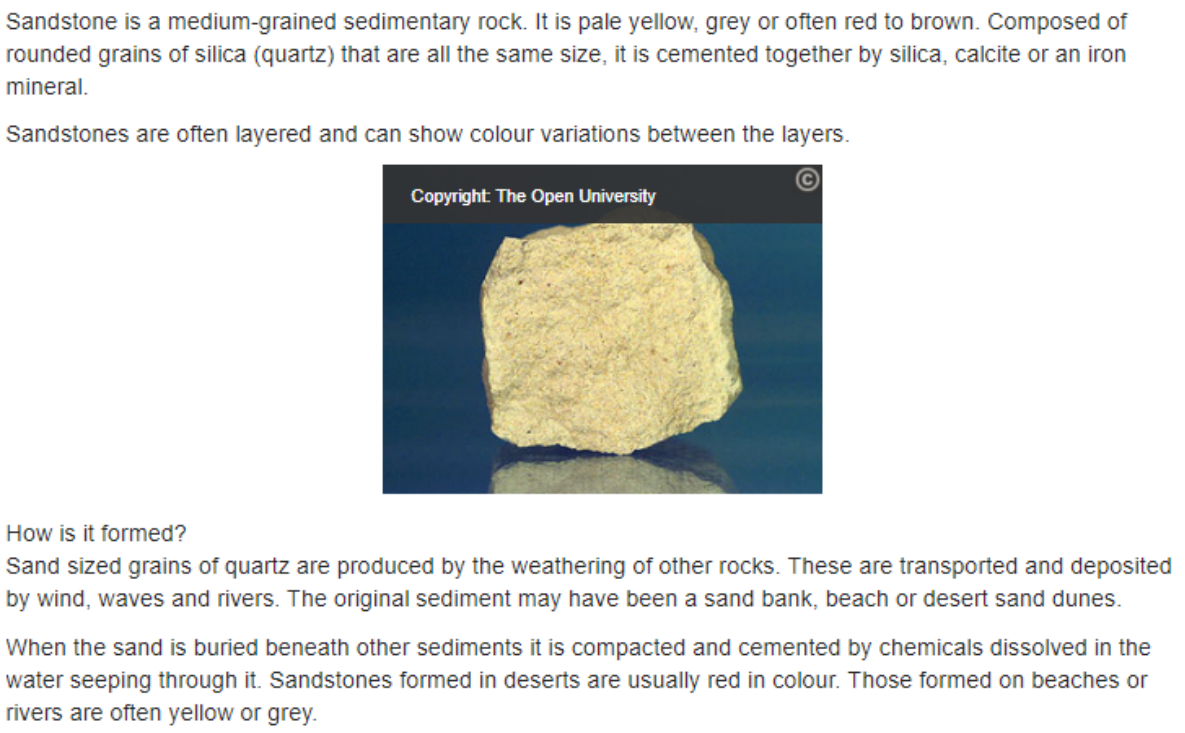


Figura 4 – Enunciado da questão *Sandstone* aplicada na *Open University*.

a *incorreta*. Esta atividade é utilizada neste capítulo como exemplo para caracterizar o material recebido pelo professor como resultado. O conteúdo completo gerado para esta atividade estão disponíveis no Capítulo B. Portanto, através dos relatórios formados, discutimos os resultados obtidos nesta questão.

Os relatórios iniciais apresentam conteúdos e dados extraídos pelo sistema. Dentre eles estão a lista de respostas dos alunos, lista de amostras requisitadas pelo sistema, lista da frequência das principais características, descrição do particionamento de treino e teste e a descrição das *features* extraídas dos documentos. Em geral, esse *feedback* é um conteúdo explicativo, apresentando o que o sistema interpretou em cada documento e algumas ações básicas tomadas durante os processos como a vetorização e contagem da frequência das características. As Tabelas 3 e 4 apresentam descrições do processamento das atividades sobre a extração de informações e particionamento de dados.

Na Tabela 3 descrevemos a forma de divisão das respostas para anotação do profes-

Dataset			Amostras
sandstone : answers.csv			1897
Treino (Un.)	Treino (%)	Teste (Un.)	Teste (%)
569	29.99	1328	70.01

Tabela 3 – Particionamento em amostras para treino e teste dos classificadores na atividade exemplo *Sandstone*.

sor e avaliação do sistema. Por outro lado, na Tabela 4 encontramos características sobre a extração de informação da atividade. Nela descrevemos o total de características encontrados e os padrões de resposta, tamanho máximo, mínimo e médio segundo quantidade de palavras e caracteres. Associada as listas de respostas, *clusters* e amostras procuramos representar o conteúdo identificado pelo sistema e, consequentemente, explicar a etapa inicial do processamento com o *p*Nota.

Outro relatório inclui a descrição do modelo avaliativo de acordo com cada um dos algoritmos testados. Independente do formato aplicado, o uso das métricas, matrizes de confusão e a comparação do desempenho de cada algoritmo ilustra a entrega de resultados do sistema. Assim, conforme os resultados finais apresentados ao fim do processo avaliativo, podemos acompanhar também a capacidade do modelo de avaliação automática gerado pelo sistema. Em diferentes perspectivas, podemos vislumbrar a capacidade de cada algoritmo avaliador. Entretanto, como principal interessado no ajuste avaliativo, destacamos a relevância do *p*Nota ilustrar a efetividade do processo avaliativo diretamente ao professor. Assim, para além de mensurar a qualidade como classificador, alinhamos os objetivos do sistema com a expectativa de resultados do professor (NASCIMENTO; KAUARK; MOURA, 2020). Para auxiliar a interpretação, dividimos em três categorias a avaliação automática sob a ótica do professor:

- Intervalo de 75 - 100%: Nível Avançado;
- Intervalo de 36 - 75%: Nível Adequado;
- Intervalo de 0 - 35%: Nível Insuficiente.

Em nível *Insuficiente* a relação entre as notas finais divulgadas pelo professor e a predição do teste para o conjunto tem desempenho baixo. Em nível *Adequado* as notas apresentam aprendizado das notas e modelos avaliativos alinhados com o professor. Por fim, em nível *Avançado*, o desempenho do classificador automático para as demais notas foi similar ao humano, identificando bem o método avaliativo do professor. É importante também descrevermos o detalhe que cada uma das métricas utilizadas pelo sistema sob a ótica do professor (NASCIMENTO; KAUARK; MOURA, 2020):

Dataset						Amostras
sandstone : answers.csv						1897
	Palavras			Caracteres		
Características	Max.	Min.	Média	Max.	Min.	Média
954	20	1	12.7	166	3	77.92

Tabela 4 – Características das respostas encontradas na atividade exemplo *Sandstone*.

- ACC: apresenta ao professor a quantidade percentual de respostas que foi avaliada de forma equivalente em relação a sua expectativa avaliativa. Através da acurácia identificamos o percentual de verdadeiros positivos, ou seja, alunos que receberam notas iguais para ambos os avaliadores, em relação ao total de respostas avaliadas. Em uma outra perspectiva podemos também considerar que, com o uso do sistema, é o percentual de notas não alteradas pelo professor após a avaliação automática.

- PRE: apresenta ao professor o percentual de acertos do total de amostras da avaliação automática em relação a sua expectativa avaliativa. A precisão exibe a quantidade de notas avaliadas de forma equivalente (verdadeiros positivos) em relação ao índice de notas que incorretamente avaliadas com o mesmo nível de nota (falsos positivos). Em geral, essa métrica determina ao professor o desempenho médio do sistema na avaliação de cada uma das notas atribuídas.

- REC: apresenta ao professor, em percentual, a capacidade do sistema em discernir cada um dos níveis de nota. A revocação é um indicativo da quantidade de notas avaliadas corretamente (verdadeiros positivos) em relação aos itens que foram categorizados de forma incoerente em outras categorias (falsos negativos). Portanto, é uma métrica que determina ao professor a capacidade média do modelo avaliativo segundo os modelos de cada nota.

- F1: é uma ponderação entre as métricas PRE e REC. Indica ao professor, em âmbito geral, a capacidade dos modelos avaliativos na identificação dos padrões de nota. Deste modo, é uma métrica que leva em conta os erros (falsos positivos e falsos negativos) do sistema entre as categorias em relação a coerência do modelo adquirido para os níveis de nota (verdadeiros positivos e verdadeiros negativos).

- MAE: apresenta ao professor o nível de erro do sistema. O erro médio absoluto determina para o conjunto de notas qual foi a diferença entre as notas atribuídas pelo professor e pelo avaliador automático. A resultante é a variação média entre as notas do professor e do sistema.

- MSE: apresenta ao professor o nível de erro do sistema com penalização. O erro médio quadrático determina a diferença entre as notas do professor e do sistema. Entretanto, a diferença quadrática entre notas atribui maior peso aos modelos avaliativos que apresentam discrepâncias maiores entre notas. A resultante, portanto, é a variação média entre notas do professor e do sistema após a penalização.

Em um contexto geral, os níveis e as métricas representam uma associação entre a consistência de cada avaliação dada pelo sistema automático e a representatividade do modelo de notas. Assim, podemos interpretar os erros sob duas perspectivas determinantes para a melhoria avaliativa e o bom uso do sistema. A primeira é a concordância entre a

correção e as componentes textuais das respostas analisadas pelo *p*Nota. E a segunda é a capacidade de montar modelos com a quantidade de respostas por cada nível de nota. O modelo formado, portanto, precisa de conteúdos coincidentes em relação aos padrões linguísticos e os modelos de nota para demonstrar o aprendizado e realizar uma avaliação efetiva. Assim, os níveis de aprendizado, *Avançado*, *Adequado* e *Insuficiente*, descrevem como cada um dos algoritmos de classificação ou regresssão foi capaz de compreender o vínculo entre as componentes textuais e o método avaliativo. Ilustramos na Figura 5, através da atividade selecionada como exemplo, os três níveis de aprendizado.

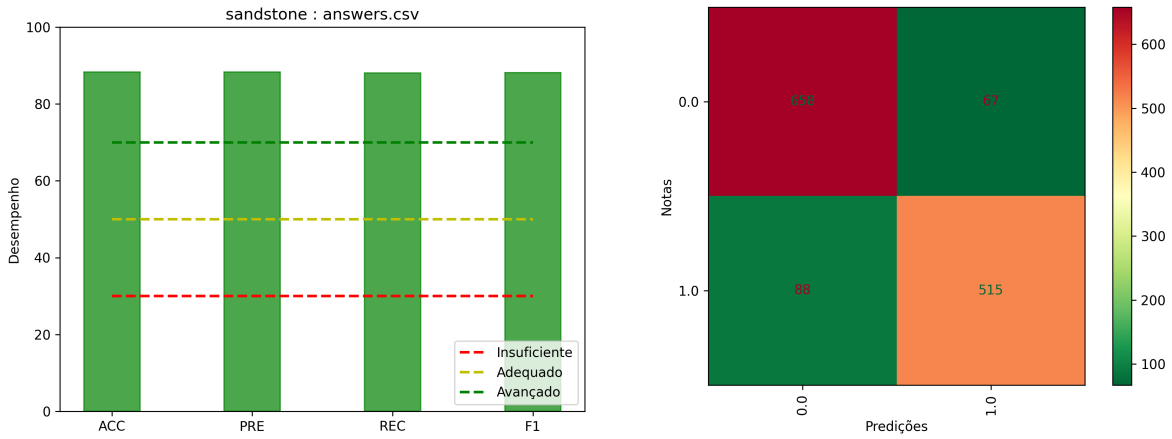


Figura 5 – Gráficos exibindo os resultados de classificação do *Random Forest* para a atividade *Sandstone* da *Open University*.

Como a Figura 5 retrata, o algoritmo RDF alcançou o índice *avançado* em todas as métricas. Ao professor isso indica que o modelo apresentou alto desempenho, com um método avaliativo muito similar ao atribuído por ele. Alinhado a isto, buscamos caracterizar o método de avaliação de forma a estabelcer para todos os participantes o desempenho obtido. Com isso, esperamos comparar o desempenho inicial com a expectativa de resultados para revisão dos métodos e identificar qualquer falha no sistema. Por conta disso, tais resultados de desempenho são dados quando a atividade encontra-se finalizada e as notas atribuídas são dadas como finais.

3.4.1 Identificação de Respostas Candidatas

Apesar do acompanhamento da dinâmica do sistema no processo avaliativo, é complexo ao sistema identificar padrões coerentes de resposta. Para isso, utilizamos o quadro de *rubrics* para representar o modelo avaliativo elaborado pelo sistema em conjunto com o professor. O quadro de *rubrics* é um modelo de caracterização do processo avaliativo conforme o modelo de resposta esperado para cada nota. Após o processo avaliativo esse processo torna-se um descritor, determinando na perspectiva dos estudantes quais foram as principais características elencadas para cada nota.

Para criação do quadro de *rubrics*, ou rúbricas, os exemplos que receberam mesma nota são considerados alinhados com uma expectativa de resposta. Sabendo que o método não utiliza ou compõe respostas candidatas, a ideia é que este processo elenque uma série de respostas para representar cada grau de nota. Para isso, utilizamos o Latent (LDA). O LDA é um método que identifica o grau de correlação do grupo de respostas de mesma avaliação e cada uma das características. A resultante é uma seleção dos 10 principais termos que compõe as respostas selecionadas. Assim, formamos o quadro de *rubrics* organizando as respostas conforme a proximidade de cada uma com as palavras selecionadas. Um exemplo de *rubrics* resultante do processo de identificação da simetria entre termos e notas é dado abaixo. A Tabela 5 define o modelo de resposta gerado pelo LDA e as respostas mais similares com o modelo.

Como podemos observar, na Tabela 5, temos os modelos de resposta para notas 0 e 1. A nota 0 em geral os alunos citam apenas a formação do arenito (*sandstone*) em regiões de deserto. Porém, em um diferencial importante, as respostas detalham a formação através da força do vento. Portanto, a função do *rubrics* é definir os padrões de resposta do sistema e, adicionalmente, criar um relatório que explique o formato avaliativo produzido em conjunto com o professor.

A ideia, portanto, é criar uma visualização do processo avaliativo de acordo com a simetria entre respostas e a avaliação observada pelo sistema. Deste modo, o uso em sala de aula relaciona diretamente cada grau de nota, as principais termos observados nas respostas e os exemplos de resposta coletados dos próprios estudantes. Por fim, como modelo descritivo da avaliação diretamente ao estudante, utilizamos a visualização de correlação entre conteúdo textual e a classificação através do Lime[3]. O Lime é uma ferramenta de visualização que descreve o processo de classificação de acordo com os padrões do conteúdo (RIBEIRO; SINGH; GUESTRIN, 2016). Para isso, é apresentado em cada resposta a correlação da resposta e suas principais componentes para cada um dos grupo de nota da avaliação. A Figura 6 apresenta essa estrutura de *feedback* diretamente ao estudante descrevendo sua avaliação.

Como a resposta da Figura 6 ilustra, podemos identificar o alinhamento das respostas com a atribuição de notas. As palavras em destaque e a correlação de cada uma com a nota caracterizam o processo de classificação automática. Deste modo, o *Lime* foi utilizado para corroborar com o critério avaliativo para que todos acompanhem os resultados encontrados, de forma similar ao *mapa de características* (SPALENZA et al., 2016b). O processo explicativo para a nota de cada resposta dá ênfase a avaliação produzida em conjunto com o professor. Para além de destacar o possível critério avaliativo em texto, possibilita a discussão dos resultados e a comparação direta entre respostas. Portanto, o material como um todo é importante para o desenvolvimento do ensino-aprendizado para

[3] Lime - https://github.com/marcotcr/lime

Sandstone	
Nota: 1.0	
Tópicos: by desert in the wind	
#	Exemplos
53	*the* sand has been transported *by wind* and *the* sandstone probably formed *in desert* conditions.
54	*the* sandstone is a *desert* sand originating *in* a *desert* environment being rounded and sorted *by wind* movement.
66	*the* sandstone originated *in the* desert. *the* shape indicates it was transported *by wind* and colour indicates oxidisation.
68	*the* sandstone transported *by wind* contains iron which when combined with oxygen gives *the* red colour found *in the* desert.
112	that *the* sandstone originated *in the desert* and were carried *by the wind*
377	*the* rock was formed *in* a *desert* and transported *by wind* oxidation of iron has occurred *in the* past.
784	*the* grains from *the* sandstone were transported *by wind in* oxidising conditions possibly *in* a *desert* situation.
1115	*the* sand was transported *by the wind* being deposited *in the* air. it comes from *desert* sand *in* oxidising conditions.
1369	formed *in* a *desert in* oxidising conditions and blown *by the wind* across *the desert* causing their well rounded shape.
1826	it originates *in the desert* as *the* grains have been transported *by wind* and *the* red colour is typical.

Sandstone	
Nota: 0.0	
Tópicos: and desert from in the	
#	Exemplos
62	*the* origin of *the* rock is *from* ancient *desert* deposits *and* has undergone chemical weathering *and* highly oxidising conditions
260	it says that *the* rock formed *in desert* conditions *and the* minerals were subjected to heavy chemical weathering *and* erosion.
349	*the* sand is originally *from the* beach *from the* wind. *the* colour is *from* iron oxide.
389	*the* sandstone comes *from* a *desert and* contains hematite *from* broken down ferromagnesian minerals *and* oxygen
669	*the* eviedence suggests that *the* sediment that *the* stone is made *from* originated *in the* dessert
670	*the* sediment that *the* stone is made *from* originated *in the* dessert *and* is wind blown *and* contains ironoxide
1037	*the* sandstone formed *from desert* sand dunes *and* was exposed to weathering *and* was not deposited *in* a slow manner.
1236	*the* sandstone resulted *from desert* sediments *the* red colour is *from* insoluble iron entering *the* rock during *the* weathering process.
1331	originating *from the desert* as *the* corners are rounded *and the* faces are pitted it has an iron oxide coating.
1737	*the* roundness of *the* grains determines *the* distance *the* sand was displaced *from the* source to *the* final deposit.

Tabela 5 – Tabela de *rubrics* para as duas notas encontradas na atividade exemplo e as respostas mais alinhadas com as palavras selecionadas pelo LDA.

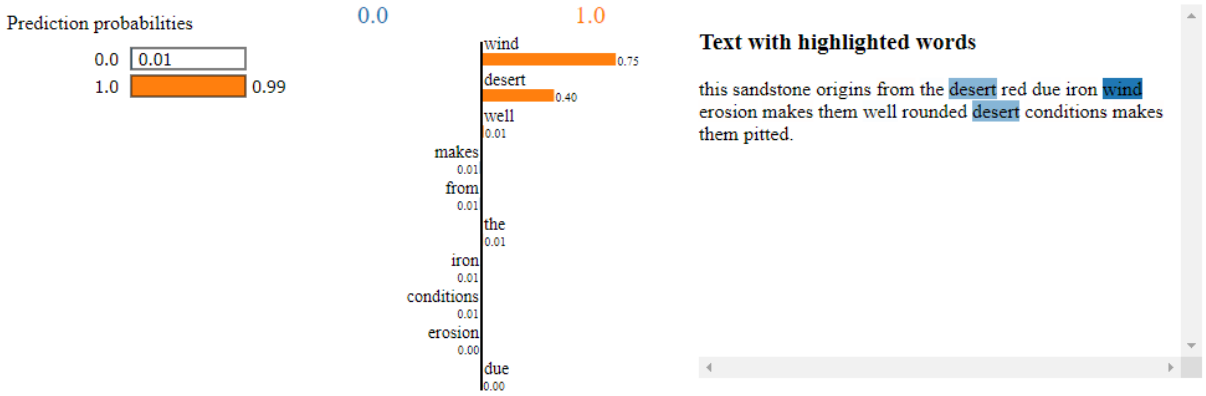


Figura 6 – Exemplo de uma resposta marcada com palavras identificadas como mais relevantes para a atividade *Sandstone* da *Open University*.

a revisão da questão e para aplicação sala de aula.

4 Experimentos e Resultados

Esse capítulo apresenta três séries de experimentos. A primeira apresenta a parte fundamental do aprendizado semi-supervisionado do sistema *p*Nota, utilizando *clusterização* para a identificação dos principais itens de resposta em cada base de dados. A segunda apresenta os métodos de classificação, a qualidade do aprendizado do sistema na predição de notas e sua adequação ao modelo esperado pelo tutor. Por fim, o terceiro módulo reflete como os modelos de resposta são formados pelo sistema e apresentados como feedback aos alunos e professores. Os experimentos foram realizados utilizando conjuntos de dados da literatura que apresentam diferentes características.

4.1 Base de Dados

Oito bases de dados foram selecionadas de acordo com a literatura, em português e inglês. Cada base de dados foi utilizada conforme as suas características. As bases de dados foram organizadas segundo o formato da nota, entre ordinais, discretas e contínua (MORETTIN; BUSSAB, 2010).

Em bases de dados com notas *ordinais* o método avaliativo do tutor é dado de forma textual e categórica. A representação do rótulo não estabelece escalas para o sistema, não sendo possível mensurar a diferenças na escala *a priori*. O modelo formado deve compreender as estruturas textuais de forma simbólica, caracterizando a essência de cada nível. Portanto, o classificador deve ser robusto para aprender a relevância das respostas pela equivalência de palavras-chave. Basicamente, é fundamental para o classificador produzir um modelo com as informações essenciais para a resposta receber tal categoria e reproduzir o modelo.

Por outro lado, outra situação acontece com bases de dados avaliados com notas contínuas. As notas *contínuas* não apresentam níveis, mas sim intervalos numéricos. As respostas recebem notas de acordo com o intervalo avaliativo. Apesar de numérico, o fato da variável não definir uma categoria que represente a divergência entre respostas dificulta o aprendizado do modelo avaliativo. Ao sistema, isso torna subjetiva a espectativa de resposta subjetivo. Assim, esse tipo de atividade é avaliada por interpolação. Nesse caso, o sistema realiza uma regressão de acordo com os pontos conhecidos, gerando a nota pela referência ao grau de similaridade para as demais respostas.

Por fim, a avaliação *discreta* numérica é a mais comum. Esse modelo favorece também os sistemas computacionais na criação da representação de resposta por categoria de nota. Ao tempo que a categoria induz a equivalência de todas as respostas ao qual foi

associada. Assim, o sistema consegue mensurar equivalência e divergências pelos indícios de proximidade entre respostas avaliadas já conhecidas para além da mesma categoria. O desafio do sistema com este tipo de nota é criar um bom modelo de classificação que aprenda essa relação dupla. Para além da categoria das respostas, o sistema passa a ter que interpretar as informações fundamentais de cada classe e a escala de divergência para as demais categorias. A Tabela 6 apresenta os detalhes de cada *dataset*, incluindo o número de questões, o total de respostas, o modelo avaliativo aplicado e a linguagem.

Base de Dados	Questões	Respostas	Modelo Avaliativo	Linguagem
SEMEVAL2013 Beetle	47	4380	ordinal	Inglês
SEMEVAL2013 SciEntsBank	143	5509	ordinal	Inglês
Kaggle ASAP-SAS	10	17043	discreto	Inglês
Powergrading	10	6980	discreto	Inglês
UK Open University	20	23790	discreto	Inglês
University of North Texas	87	2610	contínuo	Inglês
Kaggle PTASAG	15	7473	discreto	Português
Projeto Feira Literária	10	700	discreto	Português
VestUFES	5	460	contínuo	Português

Tabela 6 – Bases de dados e suas principais características.

A Tabela 6 descreve os oito *datasets* utilizados nos experimentos deste capítulo. Através das características apresentadas, sabendo que cada *dataset* contém uma quantidade regular de respotas, observamos a grande diversidade de quantidade de respostas por questão. Com questões de 30 até mais de 1800 respostas. No total, esse *corpora* apresenta um total de 337 questões e 61.268 respostas. Cada base de dados e sua descrição completa é apresentada a seguir:

4.1.1 Base de Dados *Beetle* do *SEMEVAL'2013 : Task 7 (Inglês)*

Beetle (DZIKOVSKA; NIELSEN; BREW, 2012) é um dos *datasets* utilizados durante o *International Workshop on Semantic Evaluation - SEMEVAL'2013*. O *SEMEVAL* seleciona anualmente uma série de desafios em análise semântica e apresenta no formato de competição. O *corpus Beetle* foi selecionado para a *Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge* (DZIKOVSKA et al., 2013). Portanto, a competição consistia em duas propostas. A primeira é a análise e avaliação das respostas obtidas e a segunda o reconhecimento da relação textual entre as respostas coletadas e a expectativa de resposta do professor.

Esse *dataset* consiste em uma coleção de interações entre estudantes e o sistema *Beetle II*. Beetle II é um Sistema Tutor Inteligente (STI) para aprendizado de conhecimentos

básicos em Eletricidade e Eletrônica do Ensino Médio. Os alunos foram acompanhados durante 3 a 5 horas para preparar materiais, construir e observar circuitos no simulador e interagir com o STI. Esse sistema faz questões aos alunos, avalia as respostas e envia *feedbacks* via *chat*. Na construção deste *dataset* foram acompanhados 73 estudantes voluntários da *Southeastern University* dos Estados Unidos.

Foram aplicadas questões categorizadas em dois tipos factuais e explicativas. As questões factuais requerem que o aluno nomeie diretamente determinados objetos ou propriedades. Equanto isso, as questões explicativas demandam que o aluno desenvolva a resposta em uma ou duas frases. Para a formação do *dataset* foram adicionadas apenas as atividades do segundo tipo, pois representam maior complexidade para sistemas computacionais. No total foram selecionadas 47 questões com 4380 respostas. A avaliação foi feita conforme o domínio demonstrado sobre o assunto em cinco categorias: *correct*, *partially-correct-incomplete*, *contradictory*, *irrelevant* e *non-domain*. Durante a anotação o coeficiente *Kappa* obtido foi de 69% de concordância.

4.1.2 Base de Dados *SciEntsBank* do *SEMEVAL'2013 : Task 7 (Inglês)*

O *corpus Science Entailments Bank (SciEntsBank)* (DZIKOVSKA; NIELSEN; BREW, 2012) é um dos *datasets* utilizados durante o *International Workshop on Semantic Evaluation - SEMEVAL'2013* (DZIKOVSKA et al., 2013), com foco na avaliação de sistemas conforme a sua capacidade de análise e exploração semântica da linguagem. É uma base de dados formadas pela avaliação de questões da disciplina de Ciências. Na avaliação 16 assuntos distintos são abordados entre ciências físicas, ciências da terra, ciências da vida, ciências do espaço, pensamento científico e tecnologia.

As questões são parte da *Berkeley Lawrence Hall of Science Assessing Science Knowledge (ASK)* com avaliações padronizadas de acordo com o material de apoio *Full Option Science System (FOSS)*. Participaram estudantes dos Estados Unidos de terceira a sexta série, coletando em torno de 16 mil respostas. Porém, dentre as questões de preenchimento, objetivas e discursivas, foram utilizadas apenas as discursivas, que requisitavam explicações dos alunos segundo o tema. As respostas foram graaduadas em cinco notas ordinais: *correct*, *partially-correct-incomplete*, *contradictory*, *irrelevant* e *non-domain*. Portanto, o *SciEntsBank* consiste em um conjunto com 143 questões selecionadas e 5509 respostas. No processo de avaliação foi observado o coeficiente *Kappa* com 72.8% de concordância.

4.1.3 Base de Dados do Concurso ASAP-SAS no *Kaggle (Inglês)*

A base de dados *ASAP - SAS*, *Automated Student Assessment Prize - Short Answer Scoring* é uma competição proposta pela *Hewllet Foundation* na plataforma *Kaggle* [1]. A

[1] The Hewlett Foundation - Short Answer Scoring: https://www.kaggle.com/c/asap-sas

ASAP consistiu em três fases:

- Fase 1: Demonstração em respostas longas (redações);
- Fase 2: Demonstração em respostas curtas (discursivas);
- Fase 3: Demonstração simbólica matemática/lógica (gráficos e diagramas).

O objetivo da competição foi descobrir novos sistemas de apoio ao desenvolvimento de escolas e professores. Especificamente, as três fases destacam a atividade lenta e de alto custo de avaliar manualmente testes, mesmo que com padrões bem definidos. Uma consequência disso é a redução do uso de questões discursivas nas escolas, dando preferência para as questões objetivas para evitar a sobrecarga de trabalho. Isso evidencia uma gradativa redução da capacidade dos professores em incentivar o pensamento crítico e as habilidades de escrita. Portanto, os sistemas de apoio, são uma possível solução para suportar os métodos de correção, avaliação e feedback ao conteúdo textual dos alunos.

Neste contexto, a competição apresentou 10 questões multidiciplinares, de ciências à artes. Estão distribuidas 17043 respostas de alunos dentre essas atividades. Para chegar nessa quantidade, foram selecionadas por volta de 1700 respostas dentre 3000 respostas em cada atividade. Cada resposta tem aproximadamente 50 palavras. A primeira avaliação foi dada pelo primeiro especialista como nota final e a segunda nota foi atribuída apenas para demonstrar o nível de confiança da primeira nota. A avaliação apresentada por dois especialistas apresentou concordância de 90% no coeficiente *Kappa*.

4.1.4 Base de Dados *Powergrading (Inglês)*

Elaborado através do *United States Citizenship Exam* (USCIS) em 2012, a base de dados *Powergrading* contém 10 questões e 6980 respostas (BASU; JACOBS; VANDERWENDE, 2013). Desenvolvida originalmente para destacar a possibilidade de avaliação massiva, o *dataset* selecionou 698 respostas para cada uma das questões. As respostas foram geradas com *Amazon Mechanical Turk*, serviço remoto de análise manual de conteúdo para anotação da *Amazon*. Foi coletada por um grupo de pesquisa da *Microsoft*[2] e cada questão acompanha um modelo de resposta e as respostas recebidas para cada questão. Além disso, acompanha o *dataset* outras 10 questões não avaliadas.

Foram requisitadas respostas com poucas palavras, atingindo no máximo uma ou duas sentenças. Por conta disso, os resultados são respostas muito curtas com 4 palavras em média. Em geral, por conta da convergência, vários padrões de resposta se repetem (RIORDAN et al., 2017). Com avaliações binárias, 1 para resposta correta e 0 para incorreta, cada resposta apresenta avaliações de três diferentes tutores. Apesar de alguns

[2] Powergrading: https://www.microsoft.com/en-us/download/details.aspx?id=52397

trabalhos assumirem um dos avaliadores como padrão, utilizamos como modelo de avaliação a resultante da comparação entre os três. Apesar de não ter complexidade linguística, ocorreu contradição entre os avaliadores em 470 respostas. Em valores percentuais, isso representa 7% do total de respostas do conjunto de dados.

4.1.5 Base de Dados da *UK Open University (Inglês)*

A base de dados da *UK Open University* é um conjunto de questões coletadas na disciplina de introdução à ciências, denominada *S103 - Discovering Science* (JORDAN, 2012). O foco do conjunto de atividades são abordagens em questões factuais bem concisas, não excedendo 20 palavras. Os alunos receberam as atividades através do ambiente da *Intelligent Assessment Technologies (IAT)*, o *FreeText Author*. O *FreeText Author* foi utilizado como um método de *CAA* de modo interativo e com resultado automático analizando a resposta do aluno segundo os padrões de resposta conhecidos. O sistema permitiu uma sequência de envios e apresentava comentários da resposta como *feedback* para os alunos. Dependendo da complexidade da resposta, o tempo de retorno dos resultado varia muito entre alguns poucos minutos até mais do que um dia.

Dentre as 20 questões, esse *dataset* apresenta diferentes quantidades de respostas entre 511 e 1897. A avaliação é discreta e binária, definindo cada resposta como correta ou incorreta. Não existe notas intermediárias, representando diretamente se o aluno atendeu ou não os requisitos da resposta.

4.1.6 Base de dados da *University of North Texas (Inglês)*

O *dataset* da *University of North Texas - UNT* (MOHLER; BUNESCU; MIHALCEA, 2011), conhecido como *Texas dataset* [3], é uma coleção de questões discursivas extraída no curso de Ciência da Computação. Composta por 80 atividades únicas, esse conjunto é composto por dez listas de exercícios com até sete questões e dois testes com dez questões cada. Foram aplicados em um ambiente virtual de aprendizagem durante a disciplina de Estrutura de Dados para 31 alunos. No total o *dataset* é composto por 2273 respostas de alunos dentre as 80 atividades.

A avaliação foi feita com cinco notas discretas, de 5 equivalente a resposta perfeita até 0 completamente incorreta. Foram avaliadas por dois avaliadores independentes, estudantes do curso de Ciência da Computação. Para os autores, o modelo seguido pelo sistema deve ser a resultante da média ente os avaliadores, em intervalo contínuo. Dentre as notas atribuídas, 57,7% das respostas receberam a mesma nota. Enquanto isso, um nível de diferença entre as notas representou 22,9% do total de respostas. Foi constatado também que, dentre as diferenças na avaliação, o avaliador 1 atribuia maiores notas 76%

[3] Texas Dataset: https://web.eecs.umich.edu/~mihalcea/downloads.html

das vezes.

4.1.7 Base de dados *PTASAG* no *Kaggle (Português)*

A *PTASAG - Portuguese Automatic Short Answer Grading Data* é uma base de dados brasileira apresentada por (GALHARDI et al., 2018) e disponibilizada na plataforma *Kaggle* [4]. Foi coletada pela Universidade Federal do Pampa - Unipampa em conjunto com cinco professores de biologia do Ensino Fundamental. Foram criadas 15 atividades com base no sistema Auto-Avaliador CIR. Em biologia, os tópicos abordados foram sobre o corpo humano. Cada questão acompanha uma lista de conceitos, as respostas avaliadas e as respostas candidatas criadas pelos professores como referência. Foram criadas entre duas e quatro respostas candidatas contendo entre três e seis palavras-chave.

As atividades foram aplicadas ao Ensino Fundamental para 326 estudantes de 12 a 14 anos do 8º e 9º ano. Somados a estes, também foram aplicados a 333 alunos do Ensino Médio de 14 a 17 anos. As respostas foram avalidas por 14 estudantes de uma turma do último ano, considerando uma escala de notas de 0 a 3:

- Nota 0: Majoritariamente incorreta, fora de tópico ou sem sentido;
- Nota 1: Incorreta ou incompleta mas com trechos corretos;
- Nota 2: Correta mas com importantes trechos faltantes;
- Nota 3: Majoritariamente correta apresentando os principais pontos.

No total, participaram 659 estudantes com um total de 7473 respostas. Cada uma das 15 questões apresenta entre 348 e 615 respostas. Apenas 4 questões foram avalidas por mais de um avaliador para verificar a concordância entre avaliadores. O coefficiente *Kappa* observado foi de, em média, 53.25%.

4.1.8 Base de Dados do Projeto Feira Literária das Ciências Exatas *(Português)*

É um conjunto de dados coletados durante o Projeto Feira Literária das Ciências Exatas (NASCIMENTO; KAUARK; MOURA, 2020). As questões foram obtidas durante uma Atividade Experimental Problematizada por meio de um livro paradidático, ou seja, cujo objetivo primário não é o apoio didático. O livro escolhido foi *A Formula Secreta* de David Shephard.

Conforme o livro, os professores formularam 10 atividades e ministraram para 70 alunos do 5º ano de Ensino Fundamental. Essas atividades ministradas descreviam situações práticas de Química básica. No total, o conjunto de dados conta com 10 questões,

[4] PT-ASAG-2018: https://www.kaggle.com/lucasbgalhardi/pt-asag-2018

700 respostas e suas respectivas avaliações.

4.1.9 Base de Dados do Vestibular UFES *(Português)*

A base de dados VestUFES (PISSINATI, 2014) é uma amostra das questões discursivas de Português do vestibular da UFES em 2012. A amostra selecionada contém 460 respostas divididas igualmente entre as 5 questões de língua portuguesa, também referentes a respostas dadas por 92 diferentes alunos.

Cada resposta foi avaliada por dois avaliadores. De acordo com o vestibular da universidade, os a avaliadores atribuíram notas entre 0 e 2 pontos em cada questão, totalizando 10 na soma da prova. Caso houvesse divergências de mais de 1 ponto entre as correções um terceiro avaliador era acionado para reavaliar a coerência das notas. A nota das respostas do *dataset* foram redimensionadas pelo autor para o intervalo de 0 a 10 pontos. Na nova escala, as diferenças observadas entre os avaliadores foi de, em média, 1,38 pontos com desvio padrão de 1,75.

4.2 Experimentos

A série de experimentos realizados têm como intúito demonstrar o uso do *p*Nota, em conjunto com o professor, de forma a criar modelos com bom desempenho na atribução de notas. Além disso, através dos *datasets* conhecidos na literatura, podemos comparar a proposta de análise de estruturas textuais com os principais trabalhos publicados. Em uma primeira comparação avaliaremos a qualidade dos *clusters* formados. Na sequência, apresentamos comparativos com as técnicas utilizadas na literatura de forma geral. E, por fim, verificamos o desempenho do sistema quanto ao ganho de informação, validando a capacidade de aprendizado do modelo linguístico.

4.2.1 Resultados de *Clusterização*

Em uma primeira análise, observamos os resultados dos *clusters* formados para os conjuntos de atividades através do processo de otimização (SPALENZA; PIROVANI; OLIVEIRA, 2019). Analisamos os resultados segundo três métricas de informação e o grau de similaridade entre grupos. Para isso, utilizamos as métricas ARI, NMI e CA, descritas na Seção 3.2.1. Os resultados médios obtidos para os conjuntos de dados são apresentados na Tabela 7.

A Tabela 7 demonstra que, apesar da baixa relação entre *clusters* e sua representação direta como rótulos, temos excelentes resultados quanto à classificação. Isso indica que o sistema apresenta uma capacidade de particionamento dos dados em núcleos distintos de uma mesma classe. Na prática, uma categoria é representada por mais que um *cluster*, seja

Base de Dados	ARI	NMI	CA
SEMEVAL2013 Beetle	0,0590	0,1291	0,5521
SEMEVAL2013 SciEntsBank	0,0246	0,1046	0,5747
Kaggle ASAP-SAS	-0,0123	0,0190	0,5225
Powergrading	0,1009	0,1000	0,8537
UK Open University	0,0424	0,0564	0,7604
University of North Texas	-	-	-
Kaggle PTASAG	0,0577	0,0838	0,4863
Projeto Feira Literária	0,0301	0,1088	0,5529
VestUFES	-	-	-

Tabela 7 – Bases de dados e índices qualitativos de *clusterização*.

formado por *outliers* (até 10 amostras), grupos com pares de categorias ou grupos com padrões consistentes. Um quarto tipo seriam os grupos com padrões aleatórios agrupados, em geral não desejável. Neste último caso, os resultados não contribuem muito para a passagem de informações entre as etapas de amostragem por *clusterização* e classificação. Com alinhamento direto ao método de *clusterização* proposto pelos autores, o *dataset Powergrading* apresentou alto desempenho logo de partida com nosso método. Por outro lado, com grande gradação de notas, os *datasets* do *SEMEVAL Beetle* e *SciEntsBank* indicam grande complexidade quanto a homogeneidade dos grupos. Para destacar essa distinção entre os grupos e o conteúdo, apresentamos a diferença entre *centróides* na Figura 7.

Como observamos nas imagens apresentadas na Figura 7, existe a formação de grupos muito coesos. Essa coesão é explícita pela divergência do item médio (*centróide*) os demais grupos. Como vemos através da matriz, em geral os índices de similaridade são menores que 0,25. Além disso, as atividades q2, q4, q19 e q20 apresentam respectivamente CA de 0,878, 0,805, 1,000 e 0,973. Isso demonstra a formação grupos divergentes mas com forte relação com a classe de nota. Para extração de padrões isso significa resultados de alta qualidade diretamente nos clusters formados. De certa forma este caso é incomum, ao qual esperamos lidar até com resultados de menor consistência nos *clusters*. Denominamos agrupamentos consistentes aqueles que estão diretamente alinhados com a categoria atribuída. Entretanto, o cenário encontrado neste *dataset* é o melhor possível, com padrões de resposta bem dispostos em *clusters*. O caso oposto é dado pela diferença entre *centróides* na Figura 8.

Como observamos na Figura 8, as atividades EM-16b, EM-35, EM-45c e FN-17a apresentam regiões de maior similaridade. Diferentemente do que foi notado na atividade anterior, neste temos algumas áreas coincidentes, acima de 0,35 destacadas em azul. As áreas em azul indicam um grau grande de similaridade entre centróides dos grupos e, por consequência, são compatíveis em certo nível. Por conta disso, podemos considerar que tais

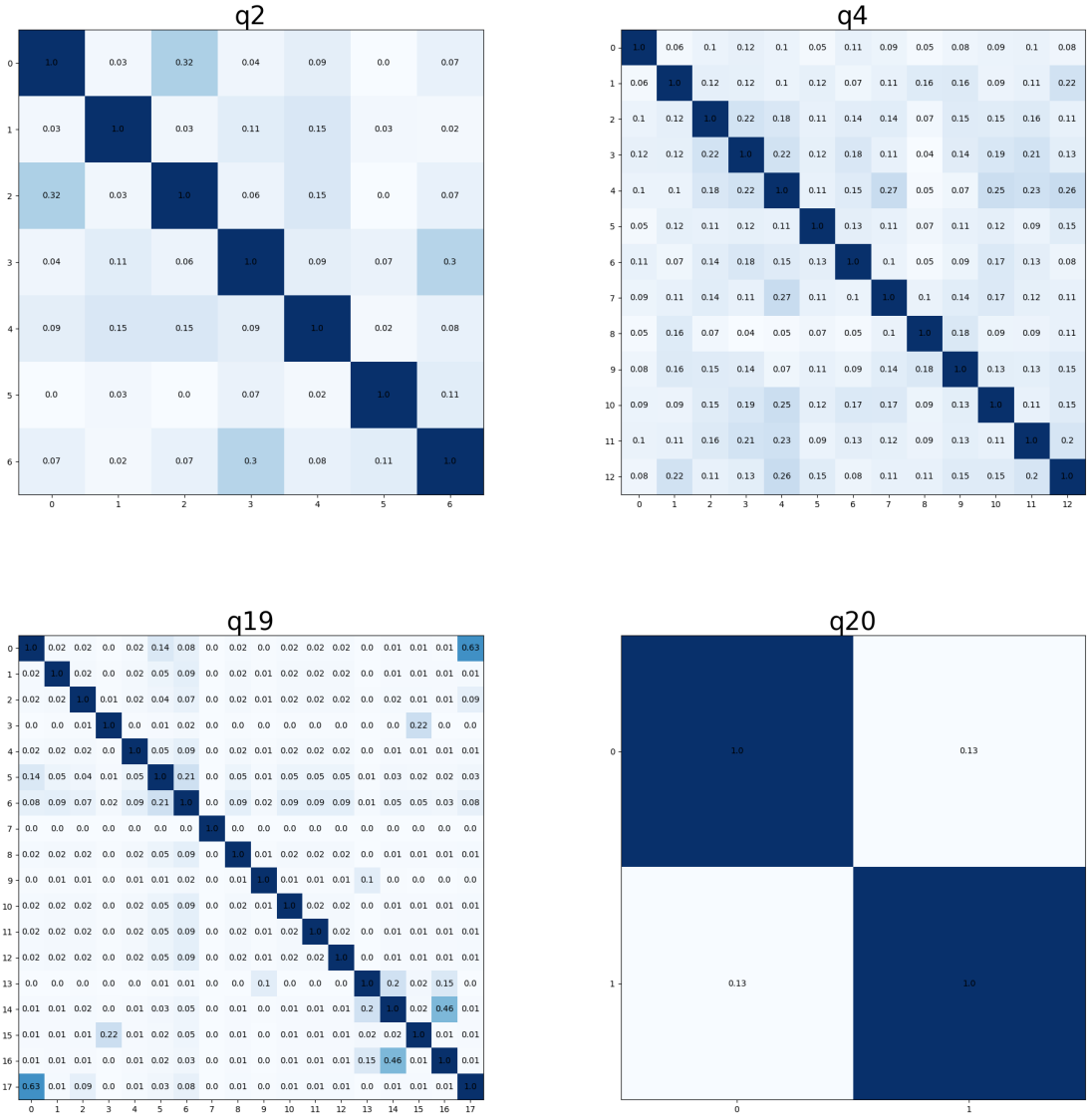


Figura 7 – Similaridade entre *centróides* para as atividades q2, q4 e q19 e q20 em *Powergrading*.

grupos podem ter diferentes perspectivas de uma mesma classe ou mesclar diferentes classes. Por consequência, em ambos os casos, torna-se maior a responsabilidade do classificador *a posteriori*, para designar padrões refinados para tais *clusters*. Por outro lado, os resultados em CA para as atividades EM-16b, EM-35, EM-45c e FN-17a são de 0,425, 0,625, 0,825 e 0,700 respectivamente. Notamos que as quatro atividades representam diferentes níveis em CA, inclusive com avaliações não coincidentes. Alinhado a isso, é fundamental que o classificador seja responsável por refinar o reconhecimento de padrões estabelecido na *clusterização*, elevando o nível dos resultados alcançados.

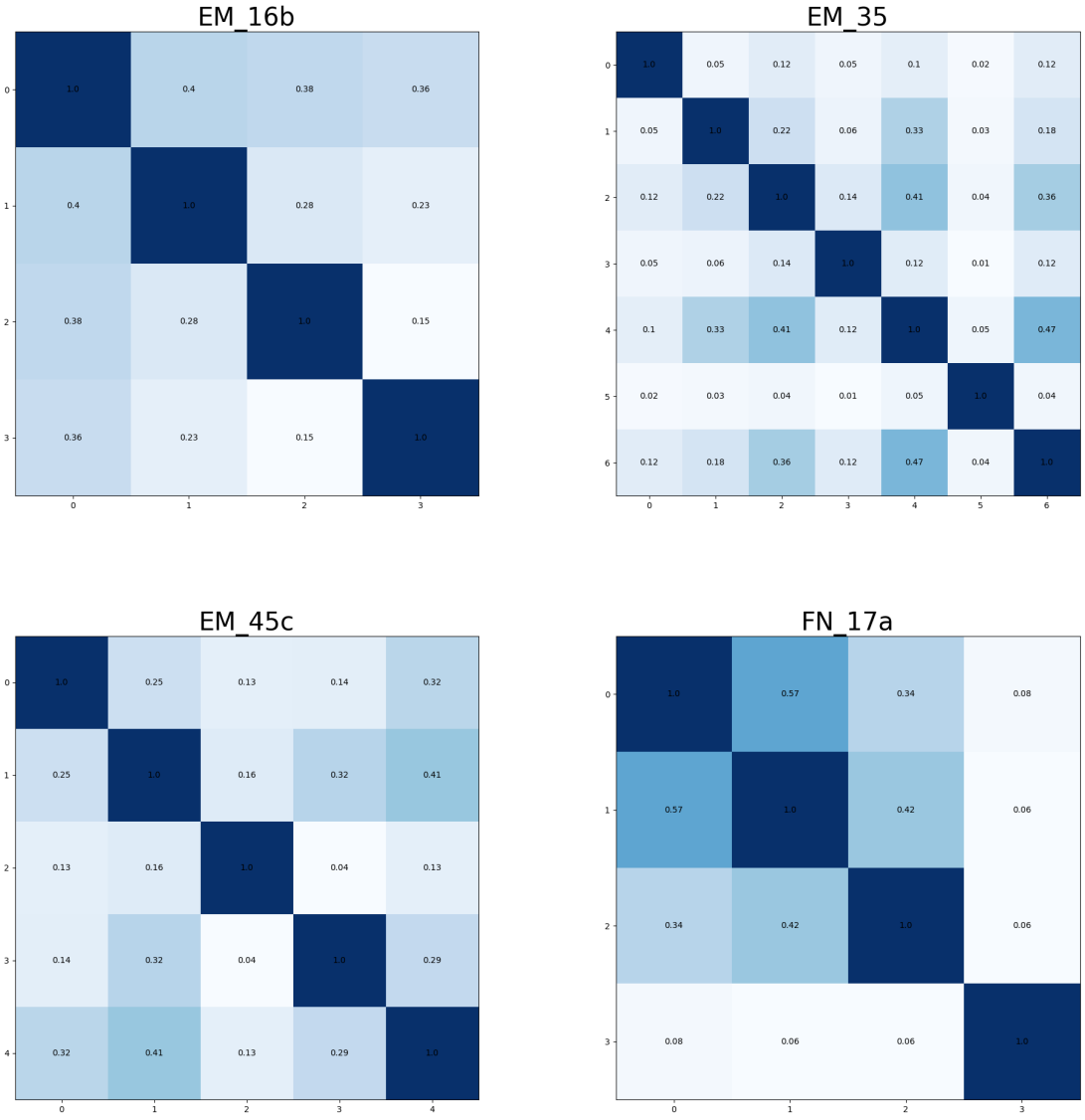


Figura 8 – Similaridade entre *centróides* para as atividades EM-16b, EM-35, EM-45c e FN-17a em *SciEntsBank*.

4.2.2 Resultados de *Classificação*

Após resultados bem sucedidos na *clusterização*, com CA por voto majoritário com médias superiores à 50%, realizamos a análise dos resultados de classificação. Tais experimentos caracterizam-se pela simetria com a maioria das publicações e desafios lançados em avaliação de respostas discursivas (BURROWS; GUREVYCH; STEIN, 2015). Nestes encontramos um particionamento de 75% de respostas selecionadas para treinamento e validação dos modelos e 25% para avaliação do desempenho das propostas. No caso deste trabalho, as partições iniciais dos *datasets* foram desconsideradas por conta da seleção de amostras de modo semi-supervisionado. Portanto, neste trabalho a amostragem dos 75% foi realizada conforme a distribuição dos itens nos *clusters* formados, como relatado em

detalhes na Seção 3.2. Portanto, foram mantidos tais percentuais como limiar superior dos sistemas. Vale ressaltar que esse processo ainda representa uma redução considerável do trabalho de correção nos mesmos 25%, apesar da maioria dos dados ser designada para avaliação do professor.

Seguindo a característica da avaliação, vamos apresentar os resultados obtidos segundo a forma avaliativa adotada pelos professores. De forma mais complexa, temos os conjuntos de dados *Beetle* e *SciEntsBank* avaliados em 5 categorias textuais (ordinais) que indicam a completude da resposta conforme a expectativa do professor. Nesse aspecto, os *datasets* do evento *SEMEVAL' 2013*, apresentam 3 níveis de desafios. O primeiro nível é a avaliação de respostas não conhecidas, selecionadas aleatóriamente no conjunto de respostas (*Unseen Answers*). O segundo nível compreende a correção de respostas em questões desconhecidas, ainda em um determinado domínio (*Unseen Questions*). E, por fim, o terceiro nível está relacionado a análise de respostas em um domínio desconhecido (*Unseen Domain*). Assim como a maioria dos sistemas SAG, o desafio que se enquadra no tópico aqui abordado é o primeiro (*Unseen Answers*), avaliando conjuntos de respostas dentro de um mesmo tópico.

Com destaque para o desbalanceamento dos dados (DZIKOVSKA et al., 2013), ambos os *datasets* foram anotados em 5 categorias: *correct*, *partially-correct-incomplete*, *contradictory*, *irrelevant* e *non-domain*. Evidenciamos ainda a complexidade, inclusive semântica, para separar as três categorias inferiores, *contradictory*, *irrelevant* e *non-domain*. Utilizando os 6 classificadores descritos na Seção 3.3.1, apresentamos os resultados obtidos na Tabela 8.

A Tabela 8 caracteriza o desempenho do sistema com os seis classificadores testados. Fica evidente que a performance do sistema é superior no *Beetle* em relação ao *SciEntsBank*. Na primeira base de dados, há um equilíbrio entre os classificadores testados, com GBC, WSD, DTR e RDF alcançando F1-ponderado entre 55% e 60%. Os classificadores KNN e SVM apresentam uma ligeira queda neste índice, com 53,73% e 52,36% respectivamente. Enquanto isso, na segunda, temos resultados consistentes do GBC e DTR, com F1-ponderado de 43% e F1-macro acima de 30%. Os demais classificadores apresentaram F1-ponderado abaixo de 40%. Deste modo, os valores em análise destacam características da formação dos conjuntos de dados. O *dataset Beetle* apresenta cerca de 84 respostas por questão. Enquanto isso, o *SciEntsBank* contém apenas 37 respostas em média por questão, ficando evidente a diferença entre eles. Os resultados obtidos também estão ilustrados na Figura 9.

Na Figura 9, os gráficos destacam os valores obtidos nas métricas F1-macro e F1-ponderado conforme os principais resultados da literatura (DZIKOVSKA et al., 2013; RAMACHANDRAN; FOLTZ, 2015; SAHU; BHOWMICK, 2020). Conforme os trabalhos da literatura, evidenciamos a distância dos resultados obtidos em relação ao que foi

Beetle							(5 Categorias)
	Métricas						
	ACC	PRE	REC	F1(m)	F1(w)	Kappa(ln)	Kappa(qu)
DTR	56,26%	31,22%	31,27%	30,43%	55,32%	0,1026	0,0802
GBC	59,39%	**36,56%**	35,08%	34,64%	**59,28%**	**0,1841**	**0,1990**
KNN	55,79%	32,95%	35,09%	32,54%	53,73%	0,1167	0,1350
RDF	58,85%	36,53%	36,31%	**34,95%**	56,03%	0,1362	0,1402
SVM	58,05%	31,50%	35,01%	31,35%	52,36%	0,0850	0,1046
WSD	**59,52%**	35,62%	**36,47%**	34,66%	56,18%	0,1391	0,1440

SciEntsBank							(5 Categorias)
	Métricas						
	ACC	PRE	REC	F1(m)	F1(w)	Kappa(ln)	Kappa(qu)
DTR	44,45%	**31,34%**	32,54%	30,45%	43,76%	0,1678	0,1433
GBC	**45,85%**	31,23%	**33,43%**	**30,71%**	**43,96%**	**0,1720**	**0,1454**
KNN	43,22%	27,19%	31,97%	27,78%	39,17%	0,0995	0,1043
RDF	43,78%	25,91%	31,26%	26,99%	38,98%	0,0994	0,0874
SVM	41,93%	21,36%	30,27%	23,53%	34,16%	0,0361	0,0345
WSD	41,89%	25,68%	29,09%	25,73%	38,09%	0,0607	0,0309

Tabela 8 – Resultados dos seis classificadores testados nos *datasets* do *SEMEVAL' 2013*.

apresentado como o *estado da arte* neste cenário. Apesar disso, ambos os resultados são superiores ao *baseline* proposto. Nesta perspectiva, abordagens com a criação de modelos de resposta, utilizando informações adicionais sobre o tema ou expressões regulares, sobressaem com melhor desempenho mas, em geral, demandam maior esforço do professor (RAMACHANDRAN; FOLTZ, 2015; SAHU; BHOWMICK, 2020). No entanto, isto indica que tais estratégias seguem um único viés de resposta, tornando-se menos efetivas ao lidar com variações linguísticas (FILIGHERA; STEUER; RENSING, 2020).

Ainda, temos dentre as 4380 respostas do *Beetle* 1841 anotadas como *correct*, 1160 como *contradictory* e 1031 como *partially-correct-incomplete*. Por outro lado, apenas 218 foram avaliadas como *non-domain* e 130 como *irrelevant*. Considerando ainda a distribuição de classes, a situação é agravada em relação ao *SciEntsBank*. Dentre as 5509 respostas, 2241 foram dadas como *correct*, 1437 como *partially-correct-incomplete*, 1248 como *irrelevant* e 557 como *contradictory*. Só constam neste *dataset* 26 respostas anotadas como *non-domain* dentre as 143 questões. Notoriamente, são poucas amostras para algumas categorias que se destacam quando vemos que, em média, o primeiro *dataset* apresenta 93 respostas por questão enquanto o segundo apresenta apenas 38 respostas. Portando, apesar da complexidade de avaliar tal questão, os resultados são positivos, aprimorando resultados esperados conforme a distribuição de *clusters*.

Em outra perspectiva, com dados discretos, analisamos os resultados dos *datasets*

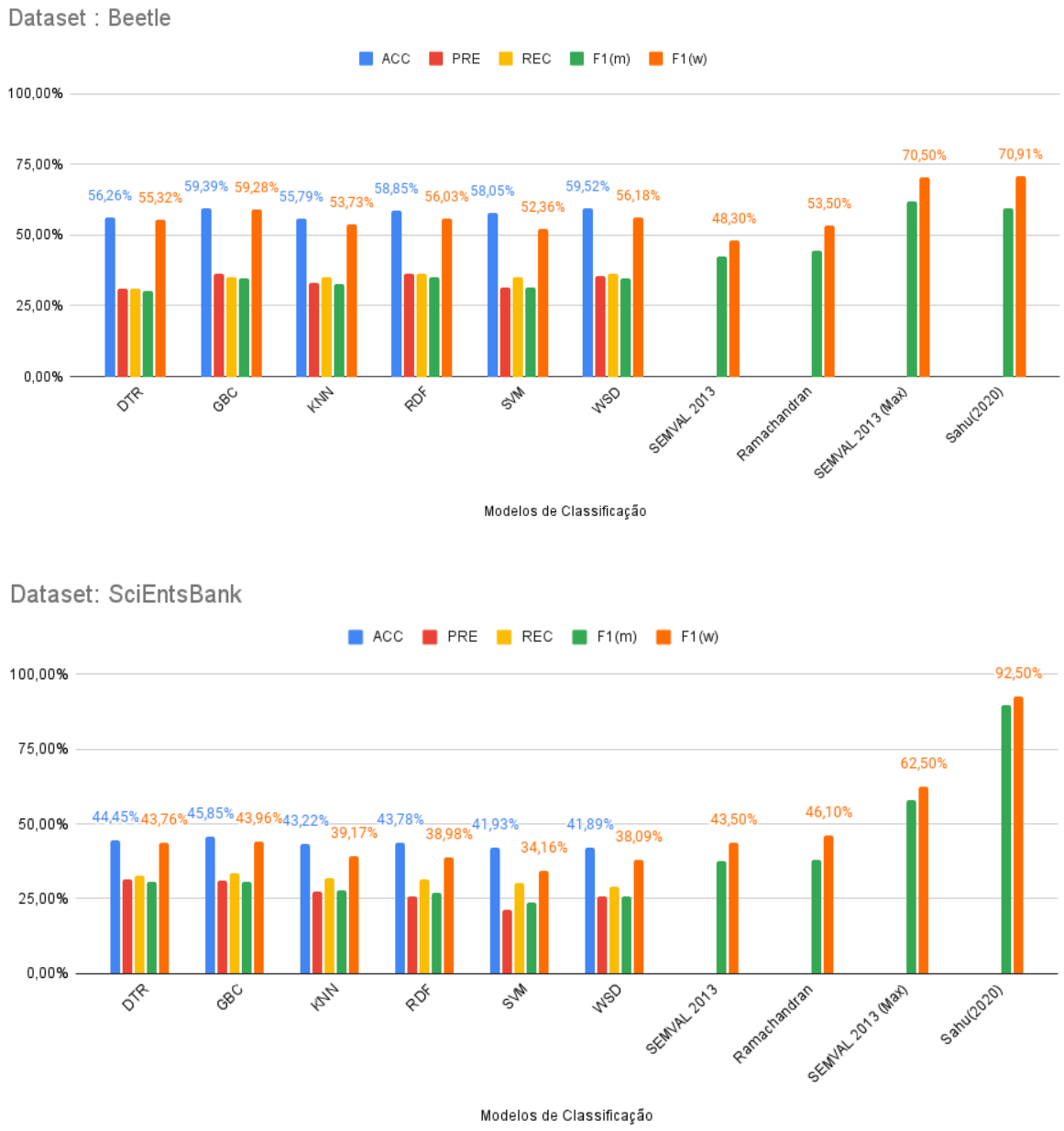


Figura 9 – Resultados obtidos nos *datasets Beetle* e *SciEntsBank* pelos classificadores em comparação com os principais encontrados na literatura.

Open University, *Powergrading* e *Kaggle ASAP-SAS*. Inicialmente abordaremos o *dataset* da *Open University*, por conta do formato de anotação. Neste conjunto de dados a cada resposta foi designada a nota 0 ou nota 1, como respostas corretas ou incorretas. Bem distinta dos *datasets Beetle* e *SciEntsBank*, este conjunto contém mais de 23 mil respostas e, em média, 1190 respostas para cada questão. Isso impacta diretamente na construção de modelos de resposta, com uma variedade de padrões para uma mesma classe, sendo possível a identificação de núcleos de resposta bem consistentes segundo a simetria da classe. Os resultados apresentados na Tabela 9 refletem justamente este aspecto.

É evidente, através dos resultados em destaque na Tabela 9, que a grande quantidade

Open University							(2 Categorias)
	Métricas						
	ACC	PRE	REC	F1(m)	F1(w)	Kappa(ln)	Kappa(qu)
DTR	95,39%	90,37%	**91,96%**	90,83%	94,86%	0,8053	0,8053
GBC	**96,15%**	**93,69%**	91,68%	**92,01%**	**95,51%**	**0,8317**	**0,8317**
KNN	89,41%	84,82%	82,60%	82,60%	88,52%	0,6264	0,6264
RDF	92,69%	89,41%	89,04%	86,86%	92,92%	0,7344	0,7344
SVM	86,46%	79,15%	77,07%	74,81%	82,98%	0,5093	0,5093
WSD	80,62%	74,53%	71,29%	70,16%	78,43%	0,3788	0,3788

Tabela 9 – Resultados de classificação para o *dataset OpenUniversity*.

de padrões de nota aumentam consideravelmente a qualidade de classificação. Com ACC de 96,15%, o algoritmo GBC captura detalhes nas respostas e formas de avaliação muito próximas do modelo do professor. Por conta disso, podemos dizer que a captura das estruturas de linguagem analisadas foram de alta qualidade. Com isso, o sistema um classificador retorna um excelente modelo de avaliação, compatível com as expectativas do professor. De acordo com as classes, os classificadores GBC, RDF e DT apresentam F1-ponderado acima de 90%. A Figura 10 apresenta os resultados obtidos neste *dataset* em relação aos descritos pelos autores em sua publicação (BUTCHER; JORDAN, 2010).

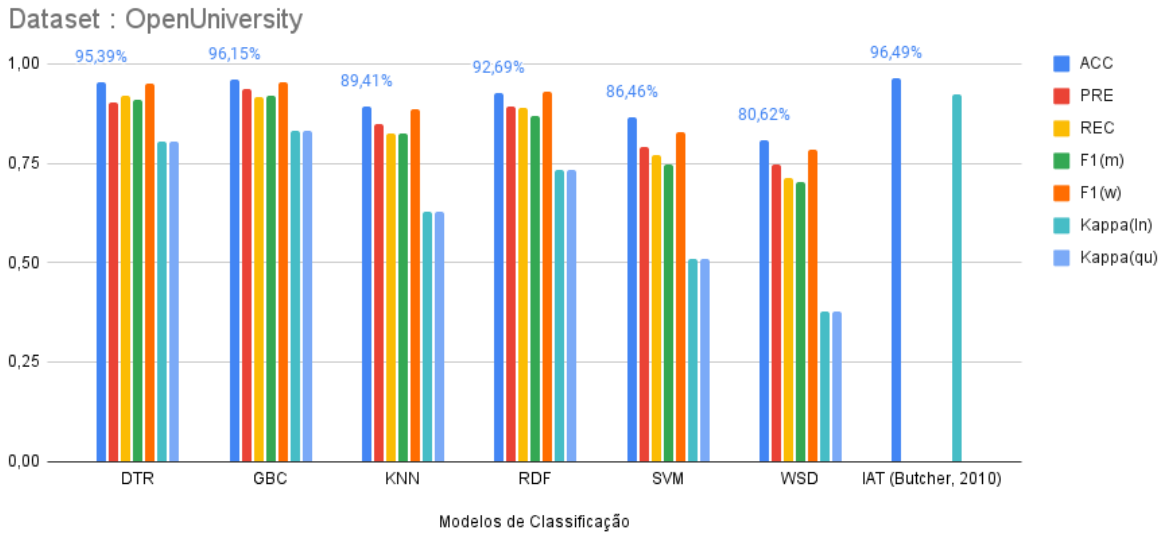


Figura 10 – Comparação do sistema dos autores do *dataset Open University* em relação ao *p*Nota.

A Figura 10 caracteriza a proximidade do modelo GBC do *p*Nota com o *IAT*, sistema aplicado na *Open University*. Como o artigo destaca, o sistema tem conhecimento sobre o conteúdo e regras de associação de respostas com o tema para a produção de *feedbacks* direcionados. Portanto, tal trabalho contém uma base de conhecimento sobre o tema para além do que é formado pelas atividades. Entretanto, apenas com os exemplos

anotados, o *p*Nota produz um modelo consistente, com desempenho equivalente aos modelos voltados ao tema.

Anotado de forma similar ao *Open University*, o *dataset Powergrading* também contém respostas avaliadas de forma binária (correto 1 ou incorreto 0). Porém, o *Powergrading* é avaliado por três avaliadores. Com as notas binárias, o objetivo é atender o resultado coincidente entre os avaliadores humanos. O desempenho obtido pelo *p*Nota neste *dataset* é caracterizado na Tabela 10.

Powergrading							(2 Categorias)
	Métricas						
	ACC	PRE	REC	F1(m)	F1(w)	Kappa(ln)	Kappa(qu)
DTR	99,37%	92,62%	91,91%	92,19%	99,27%	0,8444	0,8444
GBC	**99,49%**	93,77%	**91,97%**	**92,53%**	**99,37%**	**0,8515**	**0,8515**
KNN	99,31%	89,66%	90,00%	89,82%	98,98%	0,8000	0,8000
RDF	99,37%	**94,68%**	90,50%	90,75%	99,11%	0,8173	0,8173
SVM	99,37%	**94,68%**	90,50%	90,75%	99,11%	0,8173	0,8173
WSD	99,03%	90,39%	90,32%	90,33%	98,90%	0,8074	0,8074

Tabela 10 – Resultados de classificação para o *dataset Powergrading*.

A Tabela 10 aponta o desemepenho dos classificadores no *dataset Powergrading*, inclusive superiores a média de 90% em CA (voto majoritário). Nesse aspecto, todos os 6 classificadores apresentam ACC e F1-ponderado acima de 99%, reflexo direto da homogeneidade dos *clusters*formados na etapa anterior (BASU; JACOBS; VANDERWENDE, 2013). Em aspecto similar ao observado no *dataset Open University*, com muitas amostras os modelos das classes se tornam muito consistentes. Os resultados observados ficam evidentes na Figura 11.

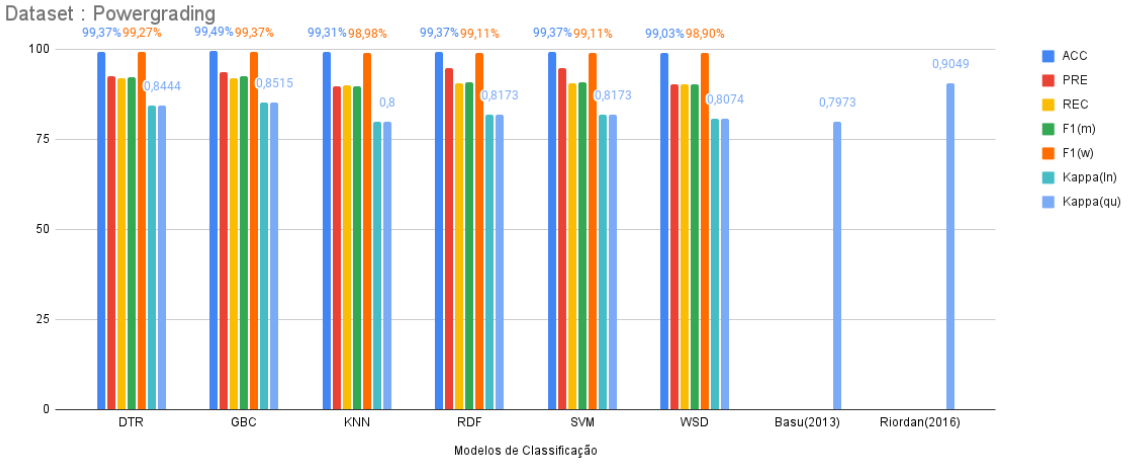


Figura 11 – Resultados dos classificadores com dados do *dataset Powergrading*.

Podemos, através da Figura 11, destacar a relação da *clusterização* na amostragem

para o classificador. Apresentando textos restritos ao tema, sucintos e em *clusters* homogêneos, são formados classificadores com alto desempenho. Deste modo, com modelos bem definidos nas etapas de *clusterização* e classificação produzimos avaliadores SAG muito similares ao tutor. Por outro lado, em alguns *datasets* a diversidade textual, a característica das notas atribuídas, os modelos linguísticos e a expectativa de resposta influenciam muito na capacidade do avaliador automático. Todas estas características são evidenciadas com conteúdos mais complexos, como a base de dados da competição do *Kaggle ASAP-SAS*.

Em outra perspectiva, temos o *dataset* da *University of North Texas* da avaliação com poucos dados. Neste conjunto, cada questão contém 30 respostas. Todas as perguntas são referentes a disciplina de Estrutura de Dados do curso de Ciência da Computação. Assim, como os demais procedimentos, sem usar respostas candidatas, utilizamos a amostragem através dos grupos resultantes da *clusterização*. Foi avaliada a atribuição de notas para três anotações do conjunto de dados: *Avaliador1*, *Avaliador2* e a *Média*. Os resultados obtidos com os algoritmos de regressão LINR, LSSR, KNRG, DTRG, WSRG na atribuição de notas entre 0 e 5 são apresentados na Tabela 11.

University of North Texas (Notas 0 - 5)			
	Métricas		
	Avaliador1		
	MAE	MSE	RMSE
LINR	1,0066	**2,5069**	**1,1955**
LSSR	1,3273	3,1713	1,4712
KNRG	0,9366	2,9032	1,2557
DTRG	**0,9233**	3,7338	1,4482
WSRG	1,2832	3,0113	1,4240
	Avaliador2		
	MAE	MSE	RMSE
LINR	**0,4752**	**0,6099**	**0,6119**
LSSR	0,6502	0,8605	0,7640
KNRG	0,4917	0,7550	0,6658
DTRG	0,5121	1,2002	0,7856
WSRG	0,6523	0,8839	0,7680
	Média		
	MAE	MSE	RMSE
LINR	**0,5058**	**0,5476**	**0,6199**
LSSR	0,7299	0,8464	0,8170
KNRG	0,5055	0,6804	0,6765
DTRG	0,5811	1,1244	0,8372
WSRG	0,7024	0,8088	0,7920

Tabela 11 – Índices de erro para cada algoritmos de regressão resultantes de cada um dos três cenários de avaliação do *dataset* da *University of North Texas*.

A Tabela 11 detalha o desempenho de cada algoritmo de regressão para as métricas MAE, MSE e RMSE. Nos resultados, destacamos a divergência entre avaliadores, com o *Avaliador 1* sendo mais irregular na análise do sistema. Por outro lado, observando os classificadores, ressaltamos a capacidade de reconhecimento de padrões mais efetiva aos modelos do *Avaliador 2* e da *Média*. A diferença entre os melhores resultados se destacam, para o primeiro avaliador obtemos RMSE de 1,1930 pontos. Enquanto isso, para o segundo obtemos RMSE de 0,6099 pontos, uma queda de 0,5831 pontos. Neste conjunto de dados, conforme a literatura, o RMSE foi utilizado como padrão para comparações com os demais trabalhos, buscando minimizar o erro obtido de acordo com a nota média. As Figuras 12 e 13 detalham os erros encontrados para as três métricas por cada um dos algoritmos e a comparação com o RMSE dos trabalhos da literatura.

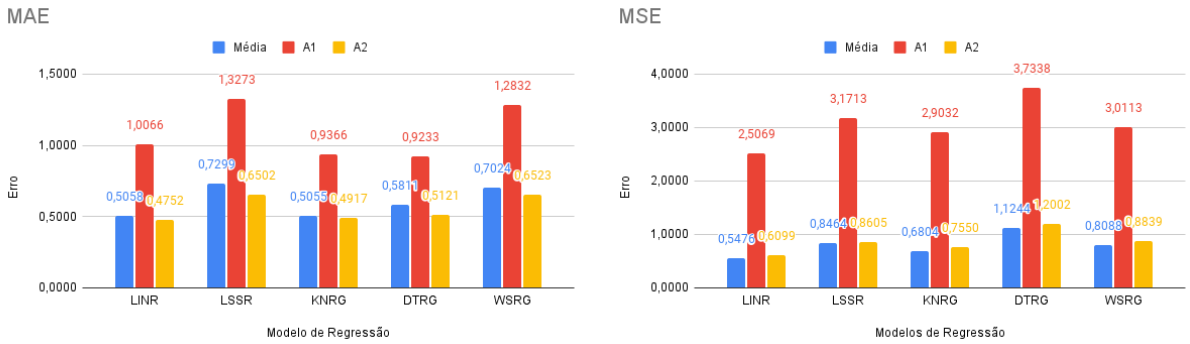


Figura 12 – Índices de MAE e MSE para os algoritmos testados em cada um dos três cenários de avaliação do *dataset* da *University of North Texas*.

Como as Figuras 12 e 13 ilustram, existe uma grande diferença entre os avaliadores. O erro observado entre humanos em RMSE é de 0,66 pontos (MOHLER; BUNESCU; MIHALCEA, 2011). Em uma comparação com demais métodos de avaliação, podemos destacar os excelentes resultados obtidos com o algoritmo de regressão linear simples LINR, como o melhor em todos os três cenários. Entretanto, dentre os algoritmos de regressão testados observamos consistência entre resultados, com pequenas variações entre os algoritmos. Enquanto na literatura o menor resultado observado é de 0,793 em RMSE, alcançamos resultados de 0,617 com LINR, seguido pelo KNRG com 0,677 pontos. Dada a avaliação com base no alinhamento com a resposta candidata elaborada pelo professor, os resultados obtidos reforçam que o uso da amostragem na produção do critério avaliativo pode ser muito efetivo. Assim, o estudo da distribuição de amostras torna o sistema conhecedor da dinâmica de avaliação realizando, neste caso, a interpolação em um espectro de notas conhecido. Esse nível, apesar da amostragem pré-estabelecida, caracteriza-se por um RMSE bem abaixo do observado até o momento na literatura (RAMACHANDRAN; CHENG; FOLTZ, 2015; KUMAR; CHAKRABARTI; ROY, 2017; SAHU; BHOWMICK, 2020).

Adicionalmente, utilizamos dados *Projeto Feira Literária* para ilustrar os resultados

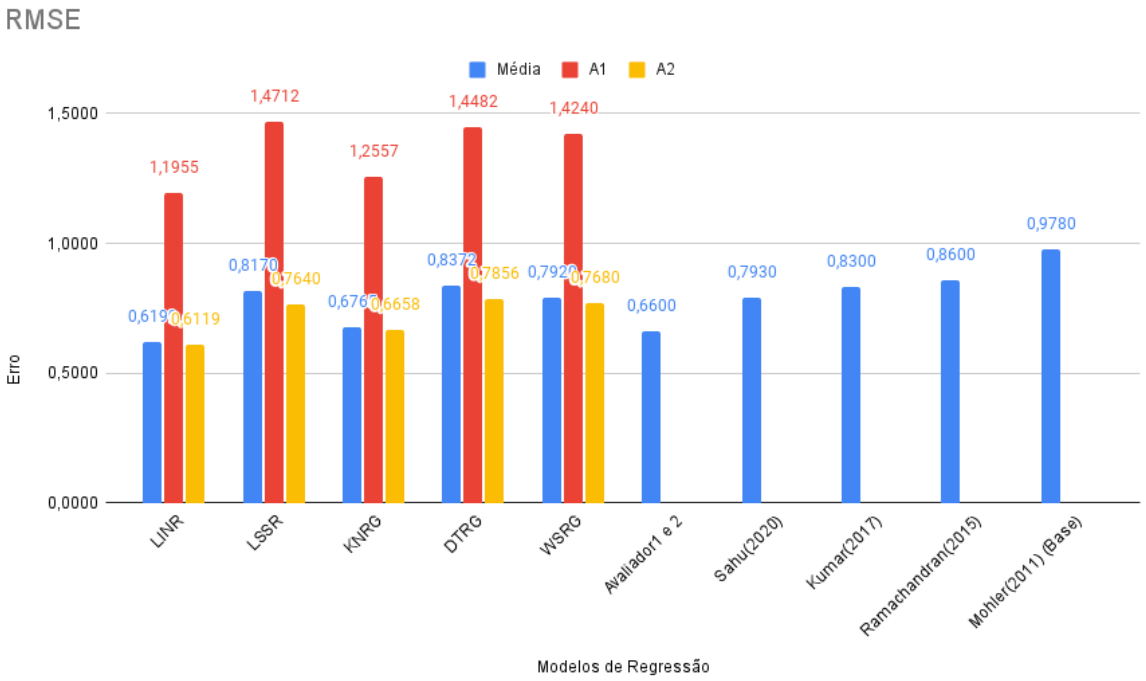


Figura 13 – Comparação entre o índice de RMSE obtidos pelo sistema e modelos propostos na literatura.

em dados nacionais. Em português, tais dados são um comparativo direto aos resultados obtidos em dados do exterior. Este dado foi coletado em conjunto com os autores para descrição da aplicação do *p*Nota e seu uso por professores (NASCIMENTO; KAUARK; MOURA, 2020). Este caracteriza-se pela presença de erros de escrita e conteúdos fora de tópico, sendo fatores avaliados negativamente pelo tutor. A Tabela 12 caracteriza o desempenho de cada um dos seis classificadores e seus resultados alcançados na avaliação das questões do projeto.

Projeto Feira Literária							(4 Categorias)
	Métricas						
	ACC	PRE	REC	F1(m)	F1(w)	Kappa(ln)	Kappa(qu)
DTR	58,59%	46,20%	43,85%	42,49%	56,75%	0,3273	0,3791
GBC	64,14%	46,21%	46,32%	43,76%	60,54%	0,3702	**0,4324**
KNN	50,00%	35,54%	34,25%	32,81%	49,64%	0,1363	0,1346
RDF	**68,18%**	**51,60%**	**51,51%**	**48,98%**	**63,56%**	**0,3842**	0,4073
SVM	57,07%	29,28%	38,52%	31,05%	47,70%	0,1121	0,1212
WSD	65,15%	48,30%	48,10%	45,25%	60,60%	0,3282	0,3543

Tabela 12 – Resultados de classificação para o *Projeto Feira Literária*.

Conforme a Tabela 12, observamos que dos seis classificadores testados, três apresentam resultados de boa qualidade. Apesar dos desafios do conjunto de dados, os algoritmos RDF, WSD e GBC apresentam resultados superiores a 60% em ACC e

F1-ponderado. Com 70 respostas e a pluralidade de estruturas textuais encontradas, destacamos a similaridade entre o desafio de correção dos datasets *Beetle* e *SciEntsBank* no ensino de ciências. Porém, podemos ressaltar como uma diferença melhores índices de PRE, REC e F1-macro. Podemos observar através da Figura 14 a proximidade destes índices com o grau de ACC alcançado.

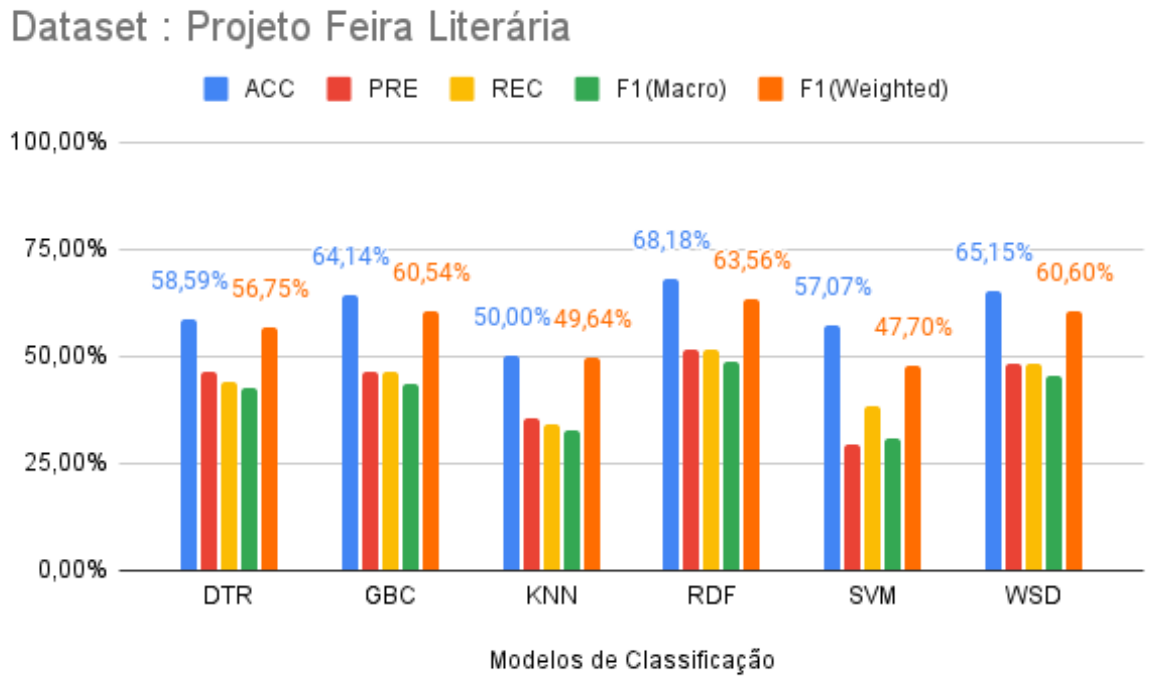


Figura 14 – Resultados alcançados para os classificadores com dados em português do *Projeto Feira Literária*.

Nessa perspectiva, a Figura 14 representa um equilíbrio dos classificadores na produção de modelos por nota e a avaliação em geral de cada categoria. Quatro classificadores, RDF, WSD, GBC e DTR, apresentam alto desempenho, enquanto KNN e SVM destoam com resultados inferiores a 50% em F1-ponderado. Portanto, os modelos avaliativos gerados demonstram complexidade mas em geral têm bons resultados.

4.2.3 Resultados de *Amostragem*

4.3 Discussão de Resultados

Os sistemas SAG têm características interessantes e de grande complexidade. Ao tempo que é um desafio lidar com diferentes tamanhos e padrões de resposta, ainda existem grande impacto o critério de avaliação. O desbalanceamento entre classes, os múltiplos padrões de resposta e a baixa quantidade de exemplos tornam complexa a criação de modelos avaliativos robustos.

De qualquer forma, por conta da dinâmica de avaliação própria de cada questão, vemos dentro de cada base de dados múltiplas perspectivas. Enquanto parte destoa com dificuldade no processo avaliativo, outras questões apresentam alto desempenho avaliativo com qualidade superior até a avaliação entre humanos. Um destaque no método avaliativo do *p*Nota quando comparado à demais abordagens da literatura é a capacidade de processamento em várias linguagens. Nesse aspecto, a adaptabilidade do sistema com a análise personalizada da estrutura linguística caracteriza-se como um diferencial das abordagens mais recentes. Em sua aplicação, o sistema demanda entre 5 minutos e 6 horas de processamento conforme o número de características e amostras do conjunto de dados, desconsiderando o processo de anotação do professor. Logo, é fundamental adequar o sistema ao uso cotidiano do tutor e seu ritmo de correção, tendo em vista a entrega rápida de resultados em sala.

Na criação do modelo linguístico, os sistemas trabalham o alinhamento entre o conjunto de respostas e as respostas candidatas. Entretanto, a proposta apresentada neste trabalho busca a evolução do modelo criado iterativamente com a avaliação do professor. Para além da análise textual o sistema prioriza a conexão entre conteúdo e critério avaliativo. Deste modo, as nuances textuais interpretadas pelo tutor durante a avaliação são destacadas com o modelo estabelecido no modelo de atribuição de notas. Superdimensionados para a avaliação de uma determinada disciplina ou tema, os modelos rígidos divergem bastante da aplicação dos sistemas SAG no cotidiano do professor. Assim, o tutor espera que o sistema seja capaz de reduzir o esforço de correção, dando suporte ao seu método assim que requisitado. Portanto, independente do cenário ao qual é aplicado, o sistema SAG deve lidar com as respostas buscando minorar a demanda de verificação do conteúdo caracterizando as demandas de ensino-aprendizagem.

Além disso, ainda é importante salientar que, apesar de serem comuns os modelos rígidos, direcionados a domínios específicos ou dependentes de regras, o modelo proposto neste trabalho foi o mesmo aplicado para todas as questões. Por conta disso, inclui-se na qualidade dos resultados obtidos a capacidade do próprio sistema na adaptação ao tema e ao modelo de avaliação. Nesta linha, poucos parâmetros são passados, tornando o sistema uma cadeia de decisões sobre o processo. A lista completa de parâmetros está no Capítulo A.

Portanto, a escolha dos modelos segundo seu alinhamento com a avaliação humana acrescenta fluidez no processo de correção. Tornamos o modelo flexível para que, cada modelo seja utilizado nas situações ao qual melhor se adequa. O nível de adequação entre o modelo e a expectativa do professor é dada através das medidas de correlação de *Pearson* ou coeficiente *Kappa*. A tendência, então, é que os métodos de classificação ou de regressão selecionados sejam os que tem desempenho superior, com o professor auditando os resultados realizando apenas pequenos ajustes.

5 Considerações Finais

O processo de avaliação de questões discursivas compreende um longo ciclo, da formulação das atividades até a análise de desempenho e redimensionamento das práticas de ensino. Portanto, é fundamental a produção de técnicas que resultem na redução do esforço do professor e a aplicação cotidiana de atividades. Em especial as que contribuem na melhoria da leitura e escrita dos estudantes em todos os níveis de instrução. Por conta disso, apresentamos neste trabalho um método semi-automático de avaliação de respostas discursivas curtas. Como destaque dos resultados obtidos com este trabalho listamos uma série de contribuições que compõe deste estudo:

- Técnica que combina *clusterização* e classificação / regressão na avaliação textual;
- Dinâmica de otimização em *clusterização* com *Gaussian Process*;
- Análise de *clusters* com amostragem por distribuição;
- Modelo vetorial complexo com múltiplas estruturas textuais;
- Identificação dos modelos de resposta, domínio e alinhamento contextual através do reconhecimento de padrões de amostras;
- Criação de formatos próprios para relatórios e *feedbacks* de questões discursivas curtas.

5.1 Conclusões

O suporte computacional do processo avaliativo compõe importante parte da integração do ensino em meio digital. Para além da avaliação automática, a avaliação em meio digital visa tornar prático e rápido o processo de avaliativo, possibilitiando sua aplicação em massa (ROMERO et al., 2010). Deste modo, o professor em um mesmo ambiente consegue interagir com todos os seus alunos e acompanhar seu desempenho na disciplina.

Por meio destas plataformas de ensino, possibilitamos o emprego de múltiplas técnicas de EDM para análise do conteúdo e, consequentemente, o aumento da capacidade avaliativa. Esse trabalho apresenta um estudo complexo de análise da estrutura textual das respostas curtas produzidas pelos estudantes. Neste, descrevemos o *p*Nota, um sistema para construção de modelos avaliativos através de interações diretas com o tutor. Integram o sistema vários módulos que, em sequência, realizam o reconhecimento de padrões

e identificação do conteúdo. Para isso compõe este processo técnicas de extração das componentes textuais, clusterização, amostragem, classificação e regressão.

Deste modo, o método semi-automático elabora uma forma de interagir com o professor para criação de modelos textuais para representar cada nota. O modelo alcançou níveis similares aos observados entre humanos na qualidade de atribuição de notas, com F1-ponderado médio de 57,84%. Destacamos que tais valores refletem que, das 200 atividades de classificação, 56 foram avaliadas com mais de 75% nesta métrica. Enquanto isso, em atividades de regressão, o RMSE médio alcançado, de 0,619 pontos, é menor do que 0,66 pontos da avaliação entre humanos. Deste modo, com esse modelo esperamos que o professor atue de forma conjunta com o sistema, corrigindo atividades e utilizando os resultados em função do desenvolvimento de seus métodos de ensino. Adicionalmente, através dos *feedbacks*, esperamos compor materiais que auxiliem várias etapas do método avaliativo, incluindo a discussão de resultados em sala.

5.2 Trabalhos Futuros

Em uma perspectiva de próximos estudos em torno do modelo proposto, despontam alguns estudos. O principal é a integração de critérios mais sofisticadas para a seleção de resultados para clusterização, classificação e regressão. Tais métodos podem ser relevantes para alcançar ainda mais qualidade na atribuição de notas, inclusive compreendendo a formação de cada conjunto de respostas. Adicionalmente, seria de grande valia, ter mais proximidade da interação do sistema sob a ótica do professor. Em especial o acompanhamento de escolas, turmas ou grupos de alunos sob a concepção das técnicas de ensino-aprendizagem. Nesse aspecto, enquadram-se os estudos detalhados da evolução dos alunos, questões aplicadas e a construção dos segmentos textuais dentro da produção textual dos estudantes.

Referências

ALEKSANDER, I.; THOMAS, W. V.; BOWDEN, P. A. Wisard - a radical step forward in image recognition. *Sensor Review*, MCB UP Ltd, v. 4, n. 3, p. 120–124, 1984. Citado na página 55.

ALMEIDA-JÚNIOR, C. R. C.; SPALENZA, M. A.; OLIVEIRA, E. de. Proposta de um Sistema de Avaliação Automática de Redações do ENEM Utilizando Técnicas de Aprendizagem de Máquina e Processamento de Linguagem Natural. In: *Computer on the Beach*. Florianópolis (SC), Brasil: Universidade do Vale do Itajaí - UNIVALI, 2017. v. 8, p. 474–483. Citado na página 35.

ARTER, J. A.; CHAPPUIS, J. *Creating & Recognizing Quality Rubrics*. 1st. ed. New York (NY), USA: Pearson Education, 2006. (Assessment Training Institute, Inc Series). Citado na página 24.

ARTSTEIN, R.; POESIO, M. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, MIT Press, v. 34, n. 4, p. 555–596, 2008. Citado 3 vezes nas páginas 27, 29 e 56.

AZAD, S. et al. Strategies for Deploying Unreliable AI Graders in High-Transparency High-Stakes Exams. In: *Proceedings of the 21st International Conference on Artificial Intelligence in Education*. Ifrane, Morocco: Springer International Publishing, 2020. (AIED' 2020, v. 21), p. 16–28. Citado 3 vezes nas páginas 27, 38 e 41.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern Information Retrieval: The Concepts and Technology behind Search*. 2nd. ed. Boston (MA), USA: Addison-Wesley Publishing Company, 2011. Citado 5 vezes nas páginas 29, 37, 39, 49 e 50.

BAILEY, S.; MEURERS, D. Diagnosing Meaning Errors in Short Answers to Reading Comprehension Questions. In: *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*. Columbus (OH), USA: Association for Computational Linguistics, 2008. (EANL '08, v. 3), p. 107–115. Citado 3 vezes nas páginas 25, 34 e 35.

BANJADE, R. et al. Evaluation Dataset (DT-Grade) and Word Weighting Approach Towards Constructed Short Answers Assessment in Tutorial Dialogue Context. In: *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*. San Diego (CA), USA: Association for Computational Linguistics, 2016. v. 11, p. 182–187. Citado na página 39.

BANJADE, R.; RUS, V.; NIRAULA, N. B. Using an Implicit Method for Coreference Resolution and Ellipsis Handling in Automatic Student Answer Assessment. In: *The Twenty-Eighth International Flairs Conference*. Hollywood (FL), USA: AAAI Press, 2015. v. 28, p. 150–155. Citado na página 36.

BARREIRA, C.; BOAVIDA, J.; ARAÚJO, N. Avaliação Formativa: Novas Formas de Ensinar e Aprender. *Revista Portuguesa de Pedagogia*, Universidade de Coimbra, v. 40, n. 3, p. 95–133, 2006. Citado na página 21.

BASU, S.; JACOBS, C.; VANDERWENDE, L. Powergrading: A Clustering Approach to Amplify Human Effort for Short Answer Grading. *Transactions of the Association for Computational Linguistics*, MIT Press, v. 1, n. 1, p. 391–402, 2013. Citado 4 vezes nas páginas 37, 39, 70 e 81.

BEZERRA, M. A. Questões Discursivas para Avaliação Escolar. *Acta Scientiarum. Language and Culture*, Universidade Estadual de Maringá, v. 30, n. 2, p. 149–157, 2008. Citado na página 34.

BIGGS, J. Assessment and Classroom Learning: A Role for Summative Assessment? *Assessment in Education: Principles, Policy & Practice*, Routledge, v. 5, n. 1, p. 103–110, 1998. Citado na página 21.

BILGIN, A. A.; ROWE, A. D.; CLARK, L. Academic Workload Implications of Assessing Student Learning in Work-Integrated Learning. *Asia-Pacific Journal of Cooperative Education*, ERIC, v. 18, n. 2, p. 167–183, 2017. Citado na página 33.

BOGARÍN, A.; CEREZO, R.; ROMERO, C. A Survey on Educational Process Mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 8, n. 1, p. e1230, 2018. Citado na página 21.

BURROWS, S.; GUREVYCH, I.; STEIN, B. The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, Springer, v. 25, n. 1, p. 60–117, 2015. Citado 10 vezes nas páginas 25, 28, 30, 33, 34, 36, 41, 42, 43 e 76.

BUTCHER, P. G.; JORDAN, S. E. A Comparison of Human and Computer Marking of Short Free-Text Student Responses. *Computers & Education*, Elsevier, v. 55, n. 2, p. 489–499, 2010. Citado 6 vezes nas páginas 26, 30, 34, 37, 41 e 80.

CALIŃSKI, T.; J., H. A Dendrite Method for Cluster Analysis. *Communications in Statistics*, Taylor & Francis, v. 3, n. 1, p. 1–27, 1974. Citado na página 51.

CAMUS, L.; FILIGHERA, A. Investigating Transformers for Automatic Short Answer Grading. In: *Proceedings of the 21st International Conference on Artificial Intelligence in Education*. Ifrane, Morocco: Springer International Publishing, 2020. (AIED' 2020, v. 21), p. 43–48. Citado na página 42.

CASIRAGHI, B.; ALMEIDA, L. S. Elaboração de um Instrumento de Avaliação do Pensamento Crítico em Estudantes Universitários. In: *Atas do V Seminário Internacional Cognição, Aprendizagem e Desempenho*. Braga, Portugal: CIEd-Universidade do Minho Portugal, 2017. v. 5, p. 30–41. Citado na página 21.

CHAKRABORTY, U. K.; ROY, S.; CHOUDHURY, S. A Fuzzy Indiscernibility Based Measure of Distance between Semantic Spaces Towards Automatic Evaluation of Free Text Answers. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, World Scientific, v. 25, n. 6, p. 987–1004, 2017. Citado 2 vezes nas páginas 27 e 41.

COHEN, J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, SAGE Publishing, v. 20, n. 1, p. 37–46, 1960. Citado na página 56.

CONDOR, A. Exploring Automatic Short Answer Grading as a Tool to Assist in Human Rating. In: *Proceedings of the 21st International Conference on Artificial Intelligence in*

Education. Ifrane, Morocco: Springer International Publishing, 2020. (AIED' 2020, v. 21), p. 74–79. Citado 3 vezes nas páginas 27, 30 e 38.

DAVIES, D. L.; BOULDIN, D. W. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 1, n. 2, p. 224–227, 1979. Citado na página 51.

DING, Y. et al. Don't Take "nswvtnvakgxpm" for an Answer - The Surprising Vulnerability of Automatic Content Scoring Systems to Adversarial Input. In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Virtual Event): International Committee on Computational Linguistics, 2020. v. 28, p. 882–892. Citado 4 vezes nas páginas 27, 30, 35 e 41.

DZIKOVSKA, M. et al. SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge. In: *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta (GE), USA: Association for Computational Linguistics, 2013. v. 7, p. 263–274. Citado 5 vezes nas páginas 27, 68, 69, 77 e 78.

DZIKOVSKA, M. O.; NIELSEN, R. D.; BREW, C. Towards Effective Tutorial Feedback for Explanation Questions: A Dataset and Baselines. In: *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montréal, Canada: Association for Computational Linguistics, 2012. v. 11, p. 200–210. Citado 3 vezes nas páginas 26, 68 e 69.

EVERITT, B. S. et al. *Cluster Analysis*. 5th. ed. Chichester, United Kingdom: John Wiley, 2011. Citado 3 vezes nas páginas 29, 37 e 51.

FERREIRA-MELLO, R. et al. Text Mining in Education. *WIREs Data Mining and Knowledge Discovery*, Wiley Online Library, v. 9, n. 6, p. e1332.1–e1332.49, 2019. Citado na página 33.

FILIGHERA, A.; STEUER, T.; RENSING, C. Fooling Automatic Short Answer Grading Systems. In: *Proceedings of the 21st International Conference on Artificial Intelligence in Education*. Ifrane, Morocco: Springer International Publishing, 2020. (AIED' 2020, v. 21), p. 177–190. Citado 7 vezes nas páginas 25, 26, 30, 35, 41, 42 e 78.

FOWLER, M. et al. Autograding "Explain in Plain English" Questions Using NLP. In: *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*. New York (NY), USA (Virtual Event): Association for Computing Machinery, 2021. (SIGCSE'21, v. 52), p. 1163–1169. Citado 2 vezes nas páginas 35 e 41.

FUNAYAMA, H. et al. Preventing critical scoring errors in short answer scoring with confidence estimation. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Online Event: Association for Computational Linguistics, 2020. v. 58, p. 237–243. Citado 5 vezes nas páginas 26, 27, 30, 35 e 40.

GALHARDI, L. B.; BRANCHER, J. D. Machine Learning Approach for Automatic Short Answer Grading: A Systematic Review. In: *Proceedings of the 16th Ibero-American*

Conference on Artificial Intelligence - IBERAMIA 2018. Trujillo, Peru: Springer International Publishing, 2018. (IBERAMIA 2018, v. 16), p. 380–391. Citado 2 vezes nas páginas 30 e 38.

GALHARDI, L. B. et al. Exploring Distinct Features for Automatic Short Answer Grading. In: *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*. São Paulo (SP), Brazil: Sociedade Brasileira de Computação, 2018. (XV ENIAC, v. 15), p. 1–12. Citado 3 vezes nas páginas 38, 42 e 72.

GHAVIDEL, H.; ZOUAQ, A.; DESMARAIS, M. Using BERT and XLNET for the Automatic Short Answer Grading Task. In: *Proceedings of the 12th International Conference on Computer Supported Education (CSEDU)*. Prague, Czechia (Virtual Event): SciTePress, 2020. (CSEDU 2020, v. 12), p. 58–67. Citado na página 40.

GOLDBERG, Y.; HIRST, G. *Neural Network Methods in Natural Language Processing*. 1st. ed. San Rafael (CA), USA: Morgan & Claypool Publishers, 2017. Citado na página 40.

GUNTHER, H.; LOPES-JÚNIOR, J. Perguntas Abertas Versus Perguntas Fechadas: Uma Comparação Empírica. *Psicologia: Teoria e Pesquisa*, Universidade de Brasília, v. 6, n. 2, p. 203–213, 2012. Citado na página 25.

HAN, J.; PEI, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. 3rd. ed. Waltham (MA), USA: Elsevier, 2011. Citado 2 vezes nas páginas 28 e 52.

HEILMAN, M.; MADNANI, N. The Impact of Training Data on Automated Short Answer Scoring Performance. In: *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver (CO), USA: Association for Computational Linguistics, 2015. v. 10, p. 81–85. Citado 2 vezes nas páginas 30 e 36.

HIGGINS, D. et al. *Is Getting the Right Answer Just About Choosing the Right Words? The Role of Syntactically-Informed Features in Short Answer Scoring*. Princeton (NJ), USA, 2014. Citado 3 vezes nas páginas 26, 30 e 40.

HORBACH, A.; PINKAL, M. Semi-supervised clustering for short answer scoring. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. (LREC 2018, v. 11), p. 4065–4071. Citado 4 vezes nas páginas 26, 27, 36 e 40.

JIMENEZ, S.; BECERRA, C.; GELBUKH, A. SOFTCARDINALITY: Hierarchical Text Overlap for Student Response Analysis. In: *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta (GA), USA: Association for Computational Linguistics, 2013. v. 7, p. 280–284. Citado 3 vezes nas páginas 36, 40 e 41.

JOHNSTONE, K. M.; ASHBAUGH, H.; WARFIELD, T. D. Effects of Repeated Practice and Contextual-Writing Experiences on College Students' Writing Skills. *Journal of Educational Psychology*, American Psychological Association, v. 94, n. 2, p. 305–315, 2002. Citado na página 33.

JORDAN, S. Student Engagement with Assessment and Feedback: Some Lessons from

Short-Answer Free-Text e-Assessment Questions. *Computers & Education*, Elsevier, v. 58, n. 2, p. 818–834, 2012. Citado 6 vezes nas páginas 26, 30, 41, 42, 59 e 71.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing*. 2nd. ed. Upper Saddle River (NJ), USA: Prentice-Hall, Inc., 2009. Citado 3 vezes nas páginas 28, 38 e 40.

KAR, S. P.; CHATTERJEE, R.; MANDAL, J. K. A novel automated assessment technique in e-learning using short answer type questions. In: *Proceedings of the 1st International Conference on Computational Intelligence, Communications, and Business Analytics*. Kolkata, India: Springer Singapore, 2017. (CICBA 2017, v. 1), p. 141–149. Citado 2 vezes nas páginas 36 e 41.

KRITHIKA, R.; NARAYANAN, J. Learning to grade short answers using machine learning techniques. In: *Proceedings of the Third International Symposium on Women in Computing and Informatics*. Kochi, India: Association for Computing Machinery, 2015. (WCI '15, v. 3), p. 262–271. Citado 2 vezes nas páginas 25 e 40.

KUMAR, S.; CHAKRABARTI, S.; ROY, S. Earth Mover's Distance Pooling over Siamese LSTMs for Automatic Short Answer Grading. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. Melbourne, Australia: AAAI Press, 2017. (IJCAI'17, v. 26), p. 2046–2052. Citado 2 vezes nas páginas 39 e 84.

KUMAR, Y. et al. Get it Scored Using AutoSAS - An Automated System for Scoring Short Answers. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Honolulu (HI), USA: AAAI Press, 2019. v. 33, p. 9662–9669. Citado 6 vezes nas páginas 27, 28, 30, 37, 39 e 41.

LANDAUER, T. K.; FOLTZ, P. W.; LAHAM, D. An Introduction to Latent Semantic Analysis. *Discourse Processes*, Routledge, v. 25, n. 2, p. 259–284, 1998. Citado na página 39.

LEFFA, V. J. Análise Automática da Resposta do Aluno em Ambiente Virtual. *Revista Brasileira de Linguística Aplicada*, SciELO, v. 3, n. 2, p. 25–40, 2003. Citado na página 25.

LIMA-FILHO, A. S. et al. *wisardpkg – A Library for WiSARD-Based Models*. Rio de Janeiro (RJ), Brazil, 2020. Citado na página 55.

LUN, J. et al. Multiple Data Augmentation Strategies for Improving Performance on Automatic Short Answer Scoring. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. New York (NY), USA: AAAI Press, 2020. v. 34, n. 09, p. 13389–13396. Citado na página 37.

MADERO, C. Secondary Teacher's Dissatisfaction with the Teaching Profession in Latin America: The Case of Brazil, Chile, and Mexico. *Teachers and Teaching*, Routledge, v. 25, n. 3, p. 358–378, 2019. Citado na página 33.

MAIMON, O.; ROKACH, L. *Data Mining and Knowledge Discovery Handbook*. 1st. ed. New York (NY), USA: Springer, 2005. Citado na página 51.

MANNING, C.; SCHUTZE, H. *Foundations of Statistical Natural Language Processing*. 1st. ed. Cambridge (MA), USA: MIT Press, 1999. Citado 2 vezes nas páginas 39 e 55.

MAO, L. et al. Validation of Automated Scoring for a Formative Assessment that Employs Scientific Argumentation. *Educational Assessment*, Routledge, v. 23, n. 2, p. 121–138, 2018. Citado na página 41.

MAQUINÉ, G. Recursos para Avaliação da Aprendizagem: Estudo Comparativo entre Ambientes Virtuais de Aprendizagem. In: *Anais do XXVI Workshop de Informática na Escola*. Natal (RN) (Online), Brasil: Sociedade Brasileira de Computação, 2020. v. 26, p. 299–308. Citado na página 21.

MARNEFFE, M.-C. et al. Universal Dependencies. *Computational Linguistics*, MIT Press, v. 47, n. 2, p. 255–308, 2021. Citado na página 48.

MARVANIYA, S. et al. Creating Scoring Rubric from Representative Student Answers for Improved Short Answer Grading. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. Torino, Italy: Association for Computing Machinery, 2018. (CIKM '18, v. 27), p. 993–1002. Citado 5 vezes nas páginas 27, 28, 31, 37 e 40.

MENINI, S. et al. Automated Short Answer Grading: A Simple Solution for a Difficult Task. In: *Proceedings of the Sixth Italian Conference on Computational Linguistics*. Bari, Italy: CEUR-WS, 2019. (CLiC-it, v. 6), p. 48.1–48.7. Citado na página 41.

MING, L. S. Reduction of Teacher Workload in a Formative Assessment Environment through use of Online Technology. In: *6th International Conference on Information Technology Based Higher Education and Training*. Santo Domingo, Dominican Republic: IEEE, 2005. v. 6, p. 18–21. Citado na página 33.

MIZUMOTO, T. et al. Analytic Score Prediction and Justification Identification in Automated Short Answer Scoring. In: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Florence, Italy: Association for Computational Linguistics, 2019. v. 14, p. 316–325. Citado 2 vezes nas páginas 31 e 37.

MOHLER, M.; BUNESCU, R.; MIHALCEA, R. Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland (OR), USA: Association for Computational Linguistics, 2011. v. 10, p. 752–762. Citado 6 vezes nas páginas 26, 38, 40, 59, 71 e 84.

MOHLER, M.; MIHALCEA, R. Text-to-text semantic similarity for automatic short answer grading. In: *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. Athens, Greece: Association for Computational Linguistics, 2009. v. 12, p. 567–575. Citado 2 vezes nas páginas 38 e 39.

MORETTIN, P. A.; BUSSAB, W. O. *Estatística Básica*. 6. ed. Pinheiros (SP), Brasil: Editora Saraiva, 2010. Citado 2 vezes nas páginas 28 e 67.

NASCIMENTO, P. V.; KAUARK, F. S.; MOURA, P. R. G. *Construindo uma Atividade Experimental Problematizada (AEP) e Avaliando Seu Nível Cognitivo de Aprendizagem Através do Software pNota no Contexto do Ensino Fundamental*. 9. ed. Vila Velha (ES), Brasil: Instituto Federal do Espírito Santo, 2020. (Série Guia Didático de Ciências/Química). Citado 3 vezes nas páginas 61, 72 e 84.

OLIVEIRA, E. et al. Combining Clustering and Classification Approaches for Reducing the Effort of Automatic Tweets Classification. In: *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval - KDIR, (IC3K 2014)*. Rome, Italy: SciTePress, 2014. (KDIR '14, v. 6), p. 465–472. Citado na página 37.

OLIVEIRA, K. L. d.; SANTOS, A. A. A. Compreensão em Leitura e Avaliação da Aprendizagem em Universitários. *Psicologia: Reflexão e Crítica*, SciELO, v. 18, n. 1, p. 118–124, 2005. Citado na página 21.

OLIVEIRA, M. G.; CIARELLI, P. M.; OLIVEIRA, E. Recommendation of Programming Activities by Multi-label Classification for a Formative Assessment of Students. *Expert Systems with Applications*, Elsevier, v. 40, n. 16, p. 6641–6651, 2013. Citado na página 33.

PADÓ, U.; PADÓ, S. Determinants of Grader Agreement: An Analysis of Multiple Short Answer Corpora. *Language Resources and Evaluation*, Springer, v. 55, n. 2, p. 1–30, 2021. Citado 6 vezes nas páginas 25, 27, 29, 30, 38 e 56.

PAIVA, R. et al. Mineração de Dados e a Gestão Inteligente da Aprendizagem: Desafios e Direcionamentos. In: *I Workshop de Desafios da Computação Aplicada á Educação (DesafIE!2012)*. Curitiba (PR), Brasil: Sociedade Brasileira de Computação, 2012. v. 1. Citado 2 vezes nas páginas 21 e 38.

PÉREZ-MARÍN, D.; PASCUAL-NIETO, I.; RODRÍGUEZ, P. Computer-Assisted Assessment of Free-Text Answers. *The Knowledge Engineering Review*, Cambridge University Press, v. 24, n. 4, p. 353–374, 2009. Citado 3 vezes nas páginas 22, 25 e 33.

PIROVANI, J. P. C. et al. Adapting NER (CRF+LG) for Many Textual Genres. In: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019)*. Bilbao, Spain: CEUR-WS, 2019. (IberLEF - SEPLN 2019, v. 35), p. 421–433. Citado na página 48.

PISSINATI, E. M. *Uma Proposta de Correção Semi-Automática de Questões Discursivas e de Visualização de Atividades para Apoio à Atuação do Docente*. Dissertação (Mestrado) — PPGI - Universidade Federal do Espírito Santo, Vitória (ES), Brasil, Set 2014. Citado na página 73.

PRIBADI, F. S. et al. Automatic Short Answer Scoring Using Words Overlapping Methods. In: *AIP Conference Proceedings*. Bandung, Indonesia: AIP Publishing LLC, 2017. v. 1818, p. 020042:1–020042:6. Citado na página 37.

RAES, A. et al. A Systematic Literature Review on Synchronous Hybrid Learning: Gaps Identified. *Learning Environments Research*, Springer, v. 23, n. 3, p. 269–290, 2020. Citado na página 21.

RAMACHANDRAN, L.; CHENG, J.; FOLTZ, P. Identifying Patterns For Short Answer Scoring Using Graph-based Lexico-Semantic Text Matching. In: *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver (CO), USA: Association for Computational Linguistics, 2015. v. 10, p. 97–106. Citado 5 vezes nas páginas 35, 36, 39, 41 e 84.

RAMACHANDRAN, L.; FOLTZ, P. Generating Reference Texts for Short Answer Scoring Using Graph-based Summarization. In: *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver (CO), USA: Association for

Computational Linguistics, 2015. v. 10, p. 207–212. Citado 4 vezes nas páginas 26, 27, 36 e 78.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco (CA), USA: Association for Computing Machinery, 2016. (KDD '16, v. 22), p. 1135–1144. Citado na página 64.

RIORDAN, B.; FLOR, M.; PUGH, R. How to Account for Mispellings: Quantifying the Benefit of Character Representations in Neural Content Scoring Models. In: *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Florence, Italy: Association for Computational Linguistics, 2019. v. 14, p. 116–126. Citado 3 vezes nas páginas 27, 38 e 41.

RIORDAN, B. et al. Investigating Neural Architectures for Short Answer Scoring. In: *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. Copenhagen, Denmark: Association for Computational Linguistics, 2017. v. 12, p. 159–168. Citado 2 vezes nas páginas 40 e 70.

ROMERO, C. et al. *Handbook of Educational Data Mining*. 1st. ed. Boca Raton (FL), USA: CRC Press, 2010. Citado 4 vezes nas páginas 28, 29, 33 e 87.

ROUSSEEUW, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, Elsevier, v. 20, n. 1, p. 53–65, 1987. Citado na página 51.

ROY, S. et al. Wisdom of Students: A Consistent Automatic Short Answer Grading Technique. In: *Proceedings of the 13th International Conference on Natural Language Processing*. Varanasi, India: NLP Association of India, 2016. v. 13, p. 178–187. Citado 2 vezes nas páginas 36 e 39.

ROY, S.; RAJKUMAR, A.; NARAHARI, Y. Selection of Automatic Short Answer Grading Techniques Using Contextual Bandits for Different Evaluation Measures. *International Journal of Advances in Engineering Sciences and Applied Mathematics*, Springer, v. 10, n. 1, p. 105–113, 2018. Citado na página 42.

SAHA, S. et al. Sentence Level or Token Level Features for Automatic Short Answer Grading?: Use Both. In: *Proceedings of the 19th International Conference on Artificial Intelligence in Education*. London, United Kingdom: Springer International Publishing, 2018. (AIED' 2018, v. 19), p. 503–517. Citado 4 vezes nas páginas 25, 27, 30 e 39.

SAHA, S. et al. *Joint Multi-Domain Learning for Automatic Short Answer Grading*. New Delhi, India, 2019. Citado 2 vezes nas páginas 27 e 42.

SAHU, A.; BHOWMICK, P. K. Feature Engineering and Ensemble-Based Approach for Improving Automatic Short-Answer Grading Performance. *IEEE Transactions on Learning Technologies*, IEEE, v. 13, n. 1, p. 77–90, 2020. Citado 5 vezes nas páginas 27, 30, 39, 78 e 84.

SAKAGUCHI, K.; HEILMAN, M.; MADNANI, N. Effective Feature Integration for Automated Short Answer Scoring. In: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies. Denver (CO), USA: Association for Computational Linguistics, 2015. v. 14, p. 1049–1054. Citado 2 vezes nas páginas 27 e 39.

SIDDIQI, R.; HARRISON, C. J. On the Automated Assessment of Short Free-Text Responses. In: *Proceedings of the 34th International Association for Educational Assessment Annual Conference*. Cambridge, United Kingdom: IAEA, 2008. (IAEA Conference, v. 34), p. 1–11. Citado na página 35.

SIEMENS, G.; BAKER, R. S. J. d. Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. In: *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*. Vancouver, Canada: Association for Computing Machinery, 2012. (LAK '12, v. 2), p. 252–254. Citado na página 33.

SPALENZA, M. A. et al. Uma Ferramenta para Mineração de Dados Educacionais: Extração de Informação em Ambientes Virtuais de Aprendizagem. In: *Computer on the Beach*. Florianópolis (SC), Brasil: Universidade do Vale do Itajaí - UNIVALI, 2018. v. 9, p. 741–750. Citado na página 44.

SPALENZA, M. A. et al. Construção de mapas de características em classes de respostas discursivas. In: *Conferência Internacional sobre Informática na Educação (TISE 2016)*. Santiago, Chile: Centro de Computación y Comunicación para la Construcción del Conocimiento (C5), 2016. (TISE 2016, v. 12), p. 630–635. Citado na página 40.

SPALENZA, M. A. et al. Uso de Mapa de Características na Avaliação de Textos Curtos nos Ambientes Virtuais de Aprendizagem. In: *XXVII Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação*. Uberlândia (MG), Brazil: Sociedade Brasileira de Computação, 2016. (SBIE 2016, v. 27), p. 1165–1174. Citado 4 vezes nas páginas 24, 27, 40 e 64.

SPALENZA, M. A. et al. Using NER + ML to Automatically Detect Fake News. In: *Proceedings of the 20th International Conference on Intelligent Systems Design and Applications*. Online Event: Springer International Publishing, 2020. (ISDA 2020, v. 20), p. 1176–1187. Citado 2 vezes nas páginas 49 e 50.

SPALENZA, M. A.; PIROVANI, J. P. C.; OLIVEIRA, E. de. Structures Discovering for Optimizing External Clustering Validation Metrics. In: *Proceedings of the 19th International Conference on Intelligent Systems Design and Applications*. Auburn (WA), USA: Springer International Publishing, 2019. (ISDA 2019, v. 19), p. 150–161. Citado 3 vezes nas páginas 51, 52 e 73.

SULTAN, M. A.; SALAZAR, C.; SUMNER, T. Fast and Easy Short Answer Grading with High Accuracy. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego (CA), USA: Association for Computational Linguistics, 2016. v. 15, p. 1070–1075. Citado na página 39.

SUNG, C. et al. Pre-Training BERT on Domain Resources for Short Answer Grading. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019. v. 9, p. 6071–6075. Citado 2 vezes nas páginas 36 e 42.

SUNG, C.; DHAMECHA, T. I.; MUKHI, N. Improving Short Answer Grading Using Transformer-Based Pre-training. In: *Proceedings of the 20th International Conference on Artificial Intelligence in Education*. Chicago (IL), USA: Springer, 2019. (AIED' 2019, v. 20), p. 469–481. Citado na página 40.

SÜZEN, N. et al. Automatic Short Answer Grading and Feedback Using Text Mining Methods. *Procedia Computer Science*, Elsevier, v. 169, n. 1, p. 726–743, 2020. Citado 3 vezes nas páginas 31, 35 e 42.

TAN, H. et al. Automatic Short Answer Grading by Encoding Student Responses via a Graph Convolutional Network. *Interactive Learning Environments*, Taylor & Francis, v. 28, n. 1, p. 1–15, 2020. Citado 2 vezes nas páginas 39 e 41.

WANG, T. et al. Inject Rubrics into Short Answer Grading System. In: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Hong Kong, China: Association for Computational Linguistics, 2019. (DeepLo 2019, v. 2), p. 175–182. Citado na página 36.

YANG, S. J. H. et al. Human-Centered Artificial Intelligence in Education: Seeing the Invisible through the Visible. *Computers and Education: Artificial Intelligence*, Elsevier, v. 2, n. 1, p. 100008, 2021. Citado 2 vezes nas páginas 38 e 41.

ZESCH, T.; HEILMAN, M.; CAHILL, A. Reducing Annotation Efforts in Supervised Short Answer Scoring. In: *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver (CO), USA: Association for Computational Linguistics, 2015. v. 10, p. 124–132. Citado 3 vezes nas páginas 26, 30 e 37.

ZESCH, T.; HORBACH, A. ESCRITO - An NLP-Enhanced Educational Scoring Toolkit. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. v. 11, p. 2310–2316. Citado 2 vezes nas páginas 39 e 42.

ZHANG, Y.; LIN, C.; CHI, M. Going Deeper: Automatic Short-Answer Grading by Combining Student and Question Models. *User Modeling and User-Adapted Interaction*, Springer, v. 30, n. 1, p. 51–80, 2020. Citado na página 40.

ZHANG, Y.; SHAH, R.; CHI, M. Deep Learning+ Student Modeling+ Clustering: A Recipe for Effective Automatic Short Answer Grading. In: *Proceedings of the 9th International Conference on Educational Data Mining*. Raleigh (NC), USA: ERIC, 2016. (EDM 2016, v. 09), p. 562–567. Citado 2 vezes nas páginas 37 e 42.

ZIAI, R.; OTT, N.; MEURERS, D. Short Answer Assessment: Establishing Links Between Research Strands. In: *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. Montreal, Canada: Association for Computational Linguistics, 2012. (NAACL HLT '12, v. 7), p. 190–200. Citado na página 41.

Apêndices

APÊNDICE A – *p*Nota

De forma descritiva, apresentamos aqui detalhes do código do *p*Nota separado em duas partes. Uma referente ao particionamento e aquisição de padrões e outra ao desenvolvimento e modeagem do critério avaliativo, ambos demonstrados através dos parâmetros e das resultantes obtidas pelo sistema. Assim, vamos separar nestas duas partes, a requisição de anotação do processo em *main_clustering.py* e a avaliação e produção de *feedbacks* em *main_classification.py*.

A.1 *main_clustering.py*

Este é o processo que envolve a *clusterização* e a seleção de itens para partições de treino e teste. Os parâmetros do processo incluem:

- (*p*Nota_module) [args] <dataset_dir>
- **dataset_dir:**

  Caminho para a pasta do dataset.
- **help: -help ou -h.**

  Descrição e ajuda.
- **data_structure: -dbstructure ou -s.**

  Modelo de arquivo(s) do dataset

  Default: moodle_files. [moodle_files, single_file, just_files]
- **data_file: -filename ou -f.**

  Arquivo de entrada de respostas.

  Default: resposta.txt
- **train_document: -trdoc ou -d.**

  Arquivo de entrada de notas.

  Default notastreino.csv
- **encoding: -encoding ou -e.**

  Padrão de caracteres.

  Default: utf-8. [utf-8, iso-8859-1]
- **n_clusters: -nclusters ou -k.**

  Valor pré-fixado de clusters ou uso de otimização.

  Default: 0 (use optimization)
- **train_method: -trmethod ou -t.**

Percentual da partição de treino.

Default: 30

- language: -language ou -l.

  Linguagem de entrada dos dados.

  Default: portuguese. [portuguese, english, spanish]

- enable_container: -container ou -c.

  Usar o modelo de container com Docker para processamento[1].

  Default: enabled. [enabled, disabled]

- enable_WordSTD: -wordmod ou -w.

  Usar stemming ou lemmatization na transformação do texto.

  Default: stemm. [stemm, lemma, none]

- enable_POS: -pos ou -p.

  Usar Part-Of-Speech Tags na transformação do texto.

  Default: enabled. [enabled, disabled]

- enable_NER: -ner ou -n.

  Usar Named Entity Recognition na transformação do texto.

  Default: enabled. [enabled, disabled]

- enable_MORPH: -morph ou -m.

  Usar análise morfológica na transformação do texto.

  Default: enabled. [enabled, disabled]

- ignore_clusters: -ignore_clusters ou -i.

  Ignorar qualquer resultado de clusterização prévia e reiniciar o processo.

  Default: disabled. [enabled, disabled]

- ngram_start: -ngram_start ou -x.

  N-gram inicial para análise de sequencias textuais.

  Default: 1

- ngram_end: -ngram_end ou -x.

  N-gram final para análise de sequencias textuais.

  Default 1

- generate_reports: -generate_reports ou -r.

  Gerar relatórios ao final do processo.

  Default: enabled. [enabled, disabled]

As resultantes desta etapa de processamento são:

- data.mtx - Matriz de vetores no formato esparso Matrix Market.

[1] Clustering-Optimization. Disponível em https://github.com/marcosspalenza/clustering_optimization

- doclist.mtx - Lista de identificadores dos documentos vetorizados.
- wordlist.txt - Lista de características extraídas e seu peso após a vetorização.
- Arquivos de *clusterização*:

  exec.csv - Cada resultado do teste de parâmetros da clusterização.

  exceptions.csv - Cada teste de parâmetros que apresentou algum tipo de exceção.

  clusters.txt - Todos os índices de cluster resultantes de clusterização dos testes completos.

  cluster{Id}.txt - particionamento do resultado selecionado em arquivos individuais de índices de cluster com o grau de similaridade encontrado.

  clstr.cfg - configurações de clusterização selecionadas pelo método de ranking.
- Arquivos de particionamento

  train - Lista de índices selecionados para anotação.

  test - Lista de índices selecionados para avaliação automática.

  minmax.mat - Lista de pares de amostras e o respectivo método de seleção adotado.
- Relatórios

  Lista de respostas.

  Distribuição de clusters.

  Lista de amostras.

  Lista de características (ocorrência máxima e mínima).

  Relatório do padrão de respostas.

  Relatório da amostragem.

  Macros.
- pNota.log - Logs do sistema.
- notastreino.csv - Lista de índices dos documentos, nota atual e requisição # de anotação.

Ao fim desta etapa a atividade contém informações de particionamento e a requisição de amostras para anotação. Além disso, parte dos relatórios são aqui gerados para apresentar, caso necessário, a etapa ao qual o processo se encontra. Inclusive, cada arquivo indica o estágio de desenvolvimento do processo e se a etapa foi corretamente executada.

A.2 *main_classification.py*

Este é o processo que envolve a classificação e a produção de *feedbacks* como resultado. Os parâmetros do processo incluem:

- (*p*Nota_module) [args] <dataset_dir>
- **dataset_dir:**

  Caminho para a pasta do dataset.

- **help: -help ou -h.**

  Descrição e ajuda.

- **data_structure: -dbstructure ou -s.**

  Modelo de arquivo(s) do dataset

  Default: moodle_files. [moodle_files, single_file, just_files]

- **data_file: -filename ou -f.**

  Arquivo de entrada de respostas.

  Default: resposta.txt

- **train_document: -trdoc ou -d.**

  Arquivo de entrada de notas.

  Default notastreino.csv

- **encoding: -encoding ou -e.**

  Padrão de caracteres.

  Default: utf-8. [utf-8, iso-8859-1]

- **language: -language ou -l.**

  Linguagem de entrada dos dados.

  Default: portuguese. [portuguese, english, spanish]

- **grade_model: -grades ou -g.**

  Padrão de atribuição de notas do ambiente.

  Default: discrete. [ordinal, discrete, continuous]

- **apply_grader: -apply_grader ou -a.**

  Permitir a avaliação para todas as amostras, inclusive os requisitos de anotação.

  Default: disabled. [enabled, disabled]

- **individual_reports: -individual_reports ou -i.**

  Habilitar a criação de feedbacks individuais.

  Default: disabled. [enabled, disabled]

As resultantes desta etapa de processamento são:

- Resultados

  {clf}.labels - Lista de predições realizadas pelo classificador.

  {clf}.performance - Qualidade obtida em cada métrica.

- Relatórios

  Marcações de resposta e correlação.

  Matriz de confusão.

  Gráfico de qualidade dos resultados.

  Quadro de rubrics.

  Regras da Decision Tree (modelo simplificado).

- `pNota.log - Logs do sistema.`
- `notastreino.csv - Lista de índices dos documentos, notas e feedbacks.`

Ao final deste processo são obtidas as notas e *feedbacks* para discussão dos resultados e revisão da avaliação com o *p*Nota. Se a etapa inicial foi efetivamente completa e as notas não foram corretamente geradas, o sistema aguarda até que tudo esteja de acordo para a produção de notas. Isso inclui esperar que todas as notas requisitadas sejam avaliadas ou que algum arquivo que não consta na lista seja criado. Portanto, nesta etapa a produção de notas envolve que tudo esteja de acordo com os pré-requisitos para execução bem como com os procedimentos anteriores.

APÊNDICE B – Exemplo

Para exemplificar o material produzido como resultados, *feedbacks* e relatórios do *p*Nota utilizamos uma questão da *Open University* denominada *Sandstone*. Essa questão foi selecionada pela diferença que a presença das palavras *wind*, *waves* e *rivers* faz segundo o critério avaliativo. As respostas que descreviam a formação da *sandstone*, ou arenito, por meio destas ações do meio sobre as particulas sedmentares foram avaliados como correto - 1. Por outro lado, não mencionar esse processo de formação resultava em nota 0. Abaixo apresentamos o enunciado da questão selecionada e a expectativa de resposta:

> Sandstone is a medium-grained sedimentary rock. It is pale yellow, grey or often red to brown. Composed of rounded grains of silica (quartz) that are all the same size, it is cemented together by silica, calcite or an iron mineral.
>
> Sandstones are often layered and can show colour variations between the layers.
>
> How is it formed?
>
> Sand sized grains of quartz are produced by the weathering of other rocks. These are transported and deposited by wind, waves and rivers. The original sediment may have been a sand bank, beach or desert sand dunes.
>
> When the sand is buried beneath other sediments it is compacted and cemented by chemicals dissolved in the water seeping through it. Sandstones formed in deserts are usually red in colour. Those formed on beaches or rivers are often yellow or grey.

B.1 Lista de Respostas

A lista de respostas é a apresentação do conteúdo do *dataset* no relatório. Nele estão os identificadores do estudante, a nota recebida (se atribuída) e a resposta dada para a questão. A Tabela 13 apresenta uma parte do relatório com 50 das respostas do conjunto.

Id	Nota	Resposta
A249457X	0.0	this sand in this rock originally was located in a desert in oxidising conditions.
A249457X	0.0	this sand in this rock originally was located in a desert in oxidising conditions.
A249457X	1.0	this sand in this rock originally was located in a hot region in windy oxidising conditions.
A2548811	1.0	it is most likely to have formed under arid desert conditions and transported by air and wind
T0632747	1.0	sedimentary oxidised desert sandstone well weathered from strong winds
Y9003030	0.0	from a desert and blown by the wind. once oxidising conditions where iron sulfide oxidised.maybe from humid conditions
A3345548	1.0	reddened suggests desert. wellsorted fine pitted suggests blown and deposited by wind. wellrounded suggests possibly a dune.
U0991840	0.0	the rock was formed in a gradually slowing current and have been exposed to oxidising conditions
U0991840	1.0	there has been wind erosion and oxidised iron and have then been rounded and deposited in a gradual current.
Y8831914	0.0	this rock layers together by the sand and grains of preexisting rocks of different size shape to form sandstone
Y8831914	0.0	sandstone is a rock sedimentary which contain rocks and sand layered together.
Y8831914	0.0	sandstone is a rock sedimentary which contain rocks and sand layered together which mean that transported by weathering.
A2415177	1.0	weathered quartz grains from a granite containing iron carried by wind in a desert.
X0816556	0.0	not deposited slowly deposition from air
X0816556	0.0	deposition from air the grains contain some iron oxid
Y6544188	0.0	the presence of reddned grains shows that sandstone derivates from sediments. sandstone is a sedimentary rock.
Y6544188	0.0	the rock is sedimentary in origin.

Continua na próxima página

Id	Nota	Resposta
Y6544188	0.0	the sandstone was formed by deposition of flowing water. the reddened grains shows that sanstone is rich in oxygen.
A2482988	0.0	it came from a higher class of diy store probably one of the quality garden centres squires rather than bq.
A2482988	0.0	rolling down river of iron oxide weathering sedimentary clastic rocks full of hyrothermal fluids
A2482988	0.0	the question is about the origins not the mode of transport as to how they got where they are.
A1424820	0.0	this rock originated in an arid desert environment that allowed physical erosion of sediments in highly oxidising conditions.
A1424820	1.0	this sandstone originated in a desert environment with wind blown physical weathering of mineral grains in highly oxidising conditions.
A2843422	0.0	the sandstone would have formed and originated in desert conditions as part of desert sands.
A2843422	0.0	the sandstone would have originated in the desert.
A2843422	1.0	the sandstone would have originated in the desert as part of a windblown dune.
Y762039X	0.0	that the sandstone originated from a desert
Y762039X	0.0	originated from a sand desert.
Y762039X	0.0	originated from dunes in a sand desert.
Y3512671	1.0	desert sandstone experiencing high oxidising conditions transported by wind
Y6523161	0.0	called a conglomerate and sedimentary in construction broken by weathering erosion transport and finally deposited.
Y6523161	0.0	were there before uncomformity occurred then weathering and erosion took place
Y6523161	1.0	grains were wind transported and were oxidised under humid conditions
A3317409	0.0	transported through water
A3317409	0.0	tha sandstone is made from sediment transported through water to its resting place.
A3317409	0.0	tha sandstone is made from sediment transported through water to its resting place.
Y8696943	0.0	it was most likely to have formed under desert conditions
Y8696943	1.0	originated as desert sandblown by wind and undergone oxidisation
Y9311424	0.0	snadstone was formed on dea bed that dried
Y9311424	0.0	sandstone was formed from sediment that entered a river and deposited
Y9311424	0.0	sandstone was formed from sediment that entered a river and deposited they were transported by water
Y8964340	1.0	that it may of came from a windy desert environment

Continua na próxima página

Id	Nota	Resposta
A2336136	0.0	the rock originated in a hot dry climate such as a desert.
A2336136	0.0	it is a desert sandstone.
A2336136	1.0	it originates in a hot dry environment weathered and transported by wind.
X8581258	1.0	these were desert grains moved by the wind finely pitted by repeated impacts and red from oxidised iron
Y1699145	1.0	they would be wind blown desert sand formed in oxidising times like the devonian period.
Y8883735	1.0	the grains have been moved by wind colliding with each other to become rounded and pitted with an iron coating.
Y8883735	1.0	very chemically weathered and moved by wind.
Y8883735	1.0	very chemically weathered and moved by wind and exposed to oxidising conditions.
		... Foram omitidos os demais resultados (1897 respostas no total).

Última página

Tabela 13 – Amostra de parte do relatório com 50 respostas extraídas da atividade exemplo.

B.2 Distribuição de *Clusters*

Com todas as respostas, o processo de *clusterização* identifica padrões de associação e agrupa amostras em subgrupos, denominados *clusters*. Apresentamos através da Tabela 14 o particionamento do conjunto de respostas e, em destaque, os itens selecionados para avaliação.

Cluster	Tamanho	Itens
0	16	68 256 **353 354** 390 513 695 **753** 754 1097 **1098** 1159 1231 1643 1843 1882
1	8	**246 247 265 266** 356 **542 1197** 1579

Continua na próxima página

Cluster	Tamanho	Itens
2	461	13 16 24 **26** 27 28 33 36 **43** 48 50 **52** 57 58 59 60 61 63 69 70 71 73 75 76 78 79 **80** 82 92 93 95 96 99 100 101 107 110 **111** 117 118 129 136 137 138 159 161 162 165 166 167 170 171 172 173 176 177 178 179 182 188 189 190 194 206 222 231 241 242 252 254 257 261 262 264 276 277 278 285 291 299 **302** 303 304 315 **316** 319 322 325 331 339 340 341 345 **347** 348 361 362 363 367 368 369 370 391 392 393 402 405 **407** 424 428 429 436 439 440 452 453 **454** 471 475 480 485 490 **491** 497 498 500 501 502 504 505 511 516 517 518 523 **524** 527 528 536 537 **538** 539 550 556 559 562 570 572 576 581 **595** 602 607 608 **609** 612 613 615 616 617 **620** 628 630 632 634 639 640 641 650 662 672 673 **675** 678 **683 685** 686 689 692 **704** 708 709 710 713 **714** 718 719 729 730 731 735 744 757 758 760 761 763 775 780 790 796 **797** 799 800 811 816 **821** 822 823 824 837 844 **845** 867 873 876 878 881 882 883 884 **890** 891 892 895 896 898 899 902 **903** 905 914 915 918 920 **922 931** 946 956 967 969 970 973 975 981 982 994 1010 1011 1014 1021 1024 1027 1032 1048 1049 1050 1051 1053 1060 1062 1063 1064 1066 1067 1071 1072 1073 1075 1078 1082 1083 1084 1085 1086 1088 1099 **1102** 1106 1108 1109 1110 **1119** 1125 1128 1129 1130 **1131** 1132 1134 1135 1137 1138 1139 1141 1142 1143 1146 1152 1153 1156 **1157** 1166 1171 **1174 1175** 1177 1187 **1188** 1192 1203 1208 **1210** 1218 1221 1224 1225 1228 1233 1234 1239 **1253** 1255 1258 1260 1268 1272 **1285** 1293 1294 **1301** 1305 1306 1319 1332 **1334** 1337 1341 1345 1353 1358 1360 1361 1366 1368 1372 1374 1380 1388 1395 1398 1400 1401 1402 **1403** 1404 1415 1430 1432 1443 1445 1450 1451 1452 1454 1457 1470 1473 1474 1475 **1478** 1481 **1483 1485** 1491 1492 1501 1502 1511 1512 1514 1523 1528 1529 1537 1538 1543 1547 1548 1555 1558 **1566 1570** 1572 1586 1587 1588 **1618** 1621 1625 1635 **1646** 1647 1651 1652 1662 1667 1670 1671 1677 1678 1683 1684 1686 1687 1688 1704 1705 1707 1716 **1717** 1721 1722 1729 1731 1736 1742 1743 1750 1753 1762 1770 1779 1781 1789 1790 1796 1797 1801 1802 1803 1806 **1808 1811 1812** 1820 1831 **1833** 1834 1835 1838 1851 **1858** 1876 1877 1880 1890 1891 **1892** 1893 **1894**
3	31	19 21 **120 122 288** 301 **379** 380 415 445 **451 478** 506 512 **603 604** 787 **802** 831 832 834 841 **916** 930 1020 1518 1576 1628 1673 1674 **1675**
4	12	**86 87** 109 342 **589** 1170 1223 1261 **1304** 1378 1696 1767
5	24	**269 310 326** 330 **371** 534 600 605 606 647 872 1195 1199 1236 1248 1265 **1510 1596** 1597 **1608** 1617 1620 1739 1862
6	50	45 62 **94 126 169 196 217** 230 320 **359 377 447 448** 487 551 **590** 614 688 691 696 **771 772** 838 863 **913 988 998** 1016 **1030 1165** 1212 1283 1331 **1342** 1357 1399 1455 **1480** 1497 1562 1564 1627 1724 **1732** 1766 **1795** 1823 1860 1871 **1886**
7	20	**5 185 410 681 769** 792 **803 828 848 864 963 1004 1022 1183 1184 1207 1591 1637** 1741 **1824**

Continua na próxima página

Cluster	Tamanho	Itens
8	300	**0 1** 3 14 20 31 32 37 38 39 42 49 56 64 67 77 81 83 89 97 112 113 119 **131** 139 146 153 160 163 164 168 **183** 195 203 208 **211** 212 **214 219** 221 224 229 232 233 235 236 **250** 258 263 270 272 282 290 **292** 293 295 300 312 321 324 328 329 332 335 338 344 350 **351** 360 364 365 372 382 383 395 396 404 **422** 427 434 441 **442 461** 464 466 474 476 479 481 499 **520** 526 529 530 531 532 555 557 561 564 567 568 574 575 577 578 579 585 **591 592** 593 598 618 624 625 631 637 638 649 658 659 660 661 665 674 676 682 690 693 705 706 715 720 721 **727** 737 738 750 762 764 765 781 **783** 793 798 807 808 809 812 814 817 825 826 827 877 887 889 897 900 904 912 921 925 929 938 940 947 948 951 961 **965** 968 971 **985** 1000 1002 1019 1026 1038 1044 **1045** 1046 1057 1059 1076 1087 1089 1090 **1101** 1113 1114 1123 1133 **1140** 1145 1158 1162 1167 1176 1178 1179 1182 1185 1186 1193 1209 1219 1222 1235 1240 1244 1249 1259 1282 1288 1295 1296 1299 1307 **1310** 1314 1315 **1316** 1321 1340 **1343** 1350 1352 1354 1359 1373 1379 1391 1406 1456 1471 1477 1482 **1484** 1488 1489 1493 1496 1503 1507 1521 **1559** 1560 1569 1577 1585 1589 1609 1610 1611 1613 1622 1623 1630 1636 1640 1653 1659 1663 1664 1668 1669 1672 1681 1685 1697 1699 1715 1719 1723 1725 1730 1746 1751 1755 1758 1761 1764 1771 1772 1775 **1776** 1782 **1785** 1786 1787 1791 1813 1821 1830 **1841** 1853 1856 1866 1879
9	38	22 91 175 **220** 259 **398 433** 449 **560** 565 **635 656** 663 **747** 779 **861** 910 932 962 1006 1120 1164 **1267** 1298 1300 1347 **1365 1431** 1440 1476 1542 **1565 1599** 1601 **1602 1676 1777** 1896
10	7	**1148 1238 1439** 1441 **1448** 1495 1740
11	37	90 181 287 336 384 521 657 664 698 **701** 702 703 **726** 743 751 **819** 840 **875** 907 911 987 1015 1052 1181 **1229 1270 1271** 1308 1309 1444 1527 1546 1594 1679 1744 1804 1867
12	250	4 **7** 8 **10 29 34 35** 40 44 53 72 74 **88 98 106 114 116 124** 128 130 **135** 197 210 **218** 234 238 239 240 **251 271** 273 **279 280** 283 **314** 318 333 334 346 366 **376** 381 385 **389 406 411 416 418 419** 425 **426** 437 438 468 **469 470 472 482 486** 488 **489** 509 **533 540** 541 **563 571 580** 587 **596 610** 611 619 **621 626 627** 636 **642** 644 **645 646** 651 **652 666** 679 **680 687** 694 **711** 716 722 **728** 739 **740 741** 742 759 766 **770 776 786** 801 804 820 **839 842 855** 856 **860** 888 **894 901** 906 **919** 926 935 941 **955** 957 **976 977** 990 **993** 1001 **1012** 1028 **1029** 1034 1040 **1041 1054** 1068 1074 **1118 1126 1144** 1147 1149 **1150** 1161 **1168** 1169 1172 **1173** 1189 1204 1206 1215 1227 1245 **1252 1256 1269** 1275 1276 1287 **1289 1302 1303** 1318 1322 **1323** 1324 **1326 1327** 1328 **1330** 1335 1336 **1338** 1339 **1346 1349** 1356 **1362 1363** 1367 **1370** 1371 **1384 1385 1386 1390 1394 1411 1417** 1418 **1422 1426** 1437 **1453** 1459 1461 **1472** 1479 **1486 1487** 1490 **1504** 1516 **1545** 1551 **1563** 1571 1573 **1575 1581** 1582 1584 **1590** 1644 1648 **1656** 1658 **1666** 1695 **1702 1708 1709** 1713 1714 1718 1745 **1747 1754 1757** 1759 1768 1788 **1798** 1815 1816 **1817** 1818 1826 1828 **1836 1837 1839 1840 1845** 1847 **1857** 1859 **1863 1872 1883 1884** 1885 1895
13	3	**156 1499 1784**
14	4	**215 846 1263** 1278
15	8	105 458 **974** 1023 1043 1421 **1424 1425**

Continua na próxima página

Cluster	Tamanho	Itens
16	25	**41** 142 202 **412 413 467** 588 **669 697** 953 **992** 995 **1107** 1121 **1273** 1274 1297 **1344 1393** 1405 1464 1469 **1567 1612** 1692
17	3	**937 1190 1434**
18	84	**51** 84 121 157 200 207 213 216 253 294 307 311 343 375 378 **394** 397 435 455 456 457 465 558 **573** 629 648 671 **677** 723 736 **745** 785 794 810 818 830 **843** 865 866 868 **870** 874 **917 980** 1007 1008 1025 1031 1035 **1091** 1095 1112 1151 1154 1198 1202 1311 **1377** 1416 1429 **1460** 1494 **1498** 1515 1519 1524 1530 1549 1593 1604 1624 **1632** 1633 1634 1680 1690 1799 1807 1814 1822 1825 1870 1878 1889
19	7	**141** 192 **933** 1241 **1355** 1435 **1438**
20	10	**226 227** 403 584 **633** 724 986 1281 **1465 1887**
21	20	191 **355** 601 712 939 954 1017 **1018** 1055 1081 1232 1442 1550 1598 **1710 1711 1712 1760** 1800 1810
22	75	23 25 144 154 **158** 174 186 193 225 296 **323** 337 401 **420** 423 430 **450** 463 483 **492** 493 503 **507** 522 **525** 543 544 670 746 **773** 782 **795** 829 **847** 871 885 **909** 942 943 964 **966** 972 1013 1093 1096 **1122** 1237 **1254 1257 1279 1280** 1286 1312 1313 1333 1387 1420 **1433** 1500 1525 1533 1556 1629 1639 1641 **1682** 1693 1698 1749 1752 **1763** 1773 1809 1829 1854
23	5	134 1033 **1375 1376 1408**
24	21	**11** 55 **132 133 184 260 473 699 777 778 789 833 853** 1111 1117 **1160 1194 1205** 1214 1508 1881
25	3	**1036 1116 1727**
26	10	459 462 **566** 1005 **1213** 1436 **1531 1848 1849** 1850
27	4	**102 508 1039** 1262
28	8	**373 547 732** 791 **813 862 1211 1700**
29	2	**858 1247**
30	2	**1615 1616**
31	9	**127 248 249 317 399** 806 1544 **1650** 1793
32	4	**755 1592 1603 1769**
33	15	**15** 17 66 103 **104** 400 **1216 1217** 1226 1290 **1291** 1389 1535 **1536** 1649
34	2	**6 1557**
35	16	201 **274** 281 408 414 583 774 854 **927 928** 997 **1047** 1277 **1583 1774 1868**
36	81	125 143 145 148 **152** 155 187 237 275 284 286 289 297 305 352 421 431 **460** 484 494 546 549 569 594 **667** 717 752 788 **835 836** 852 857 **879** 893 936 1003 1042 1058 1065 1077 **1094** 1100 1104 1105 1115 1155 **1180** 1191 1329 **1369** 1381 1382 1412 1428 1449 1458 1466 1468 1506 1509 1513 1532 1561 1578 1606 **1631** 1638 1657 1689 1691 1703 **1733** 1734 **1735** 1738 1765 1780 1794 1805 1842 1869
37	1	**999**
38	11	**386 387 388 409 432** 653 **654** 700 **1246 1383 1855**
39	2	**748 749**

Continua na próxima página

Cluster	Tamanho	Itens
40	8	**9** 228 **327 1243 1266** 1410 **1778** 1846
41	1	**495**
42	4	**869 1325 1520** 1832
43	1	**309**
44	1	**1720**
45	5	**180** 934 **1136 1397 1407**
46	9	**519** 944 **949 989 1127** 1264 1522 1541 1568
47	2	**18 960**
48	2	**622 1827**
49	6	**198 199 245 1554** 1655 1861
50	1	**859**
51	76	12 47 85 108 115 123 **149 150 151 204** 205 223 244 306 313 349 417 446 535 545 **552** 553 554 582 586 599 643 684 725 **733** 815 **880** 923 **924** 945 978 996 1009 1056 1061 1069 **1070** 1103 1163 **1196** 1200 1242 1284 1292 1320 **1364** 1413 **1414** 1419 1427 1447 1462 **1463** 1467 1526 **1553 1607** 1619 1654 **1660 1726** 1728 1748 **1756** 1783 1792 1819 **1852** 1864 **1865** 1873
52	3	**444 983 1505**
53	15	30 46 **140 147** 243 597 707 **756** 1201 1351 **1446** 1600 **1665 1694 1888**
54	43	2 54 **209** 298 357 **374** 477 510 548 623 655 734 767 **768** 784 **805** 849 **850 851** 886 **950** 958 959 979 991 1037 1079 **1080 1124** 1317 1348 1392 1539 **1540** 1580 **1595** 1626 **1642** 1645 1661 1844 **1874** 1875
55	6	443 **496 514 984 1396 1701**
56	1	**1574**
57	3	**515 1230 1614**
58	21	**65** 255 **267 268** 308 358 668 **908** 952 **1092** 1220 **1250** 1251 1409 1423 1517 1534 1552 **1605** 1706 **1737**

Última página

Tabela 14 – Clusters formados com cada uma das respostas da atividade *Sandstone*.

B.3 Lista de Amostras por *Cluster*

Com um processo amostral que combinando diferentes modos, descrevemos através de cada amostra selecionada a técnica que foi aplicada. Isso por conta de que a técnica define detalhes da representatividade de cada amostra ao *cluster* de origem, até o inicio da amostragem por distribuição. Assim, na Tabela 15 descrevemos o modo de amostragem adotado para as amostras de alguns dos *clusters* da atividade *Sandstone*.

Cluster	Id	Método	Resposta
0	353	maxsim	the sandstone was probably formed in a desert or low energy environment- reddened grains suggests oxidisation of iron took place.
0	753	maxsim	southwest of england and wales- iron oxide from southwestern united states. quartz arenites.
0	353	minsim	the sandstone was probably formed in a desert or low energy environment- reddened grains suggests oxidisation of iron took place.
0	354	minsim	the sandstone was originally a beach or similar low energy environment- reddened grains suggests oxidisation of iron took place.
0	753	minsize	southwest of england and wales- iron oxide from southwestern united states. quartz arenites.
0	354	maxsize	the sandstone was originally a beach or similar low energy environment- reddened grains suggests oxidisation of iron took place.
0	1098	silhcoeff	the sandstone is transported by wind- originated in oxygen rich atmosphere is chemically weathered and transported in high energy conditions.
1	1197	maxsim	igneous rock-formed deep within earth and cooled slowly.mainly quartz-chemically weathered- longer aeolian processed-desert enviroment
1	542	maxsim	sandstone is a sedimentary rock composed mainly of sand-size mineral or rock grains.
1	246	minsim	it would have orriganated in a dessert because there reddened grains. it would have been deposited by slow moving water.
1	247	minsim	it would have orriganated in a dessert because there reddened grains. it would have been deposited by slow moving water.
1	265	minsize	the sandstone indicates continued hot- oxidising conditions and from a sedimentary origin.
1	1197	maxsize	igneous rock-formed deep within earth and cooled slowly.mainly quartz-chemically weathered- longer aeolian processed-desert enviroment
1	266	silhcoeff	the sandstone indicates continued hot- oxidising conditions and layered. so therefore from a sedimentary origin.

Continua na próxima página

Cluster	Id	Método	Resposta
2	845	maxsim	well sorted and deposited by a current that gradually slowed by sediment slowing
2	43	maxsim	it is a desert sandstone.
2	26	minsim	that the sandstone originated from a desert
2	111	minsim	that the sandstone originated in the desert
2	302	minsize	weathered
2	845	maxsize	well sorted and deposited by a current that gradually slowed by sediment slowing
2	1858	silhcoeff	it is a sedimentary rock formed from grains from an arid desert
2	1618	silhcoeff	the sandstone grains originated in an arid desert.
2	1485	silhcoeff	the sandstone formed from ancient desert deposits.
2	1174	silhcoeff	this sandstone has actually originated from a rock in a desert.
2	1119	silhcoeff	the deposited sand originates from an ancient desert.
2	52	silhcoeff	this indicates it may be desert sandstone formed by the wind
2	538	silhcoeff	it originated from and area where iron was present in a river.
2	347	silhcoeff	the rock is from weathering which has produced a sedimentary rock.
2	1570	silhcoeff	the rock originated in the desert and was blown by the wind.
2	714	silhcoeff	it was probably transported by wind in desert conditions.
2	1301	silhcoeff	the rock is made from sand grains from the desert.
2	620	silhcoeff	the grains came from a desert sand dune.
2	524	silhcoeff	this rock appears to have its origin in desert sands.
2	454	silhcoeff	desert sand dunes sorted by wind.
2	903	silhcoeff	it indicates the origin is arid desert conditions.
2	1157	silhcoeff	it formed on land in arid conditions and was transported by wind
2	1566	silhcoeff	these indicators signify desert conditions in the past.
2	1403	silhcoeff	this is from the desert in origin it is windblown and formed on a dune.
2	1894	silhcoeff	the rock originates from ancient desert deposits.
2	407	silhcoeff	the sand was deposited as layers of an ancient desert.
2	80	silhcoeff	formed after high energy transportation under desert conditions
2	931	silhcoeff	most likely formed under highly oxidising desert conditions.
2	685	silhcoeff	arid dessert conditions sorted by wind.
2	609	silhcoeff	the rock would have originated from a desert and moved by wind
2	890	silhcoeff	it was a sedimentary rock that was formed under desert cinditions.
2	1478	silhcoeff	sedimentary rock formed in deserts and contains iron
2	1210	silhcoeff	this indicates desert conditions and windy conditions in the past
2	675	silhcoeff	formed in desert conditions in an oxygen rich atmosphere

Continua na próxima página

Cluster	Id	Método	Resposta
2	704	silhcoeff	formed in desert conditions in oxygen rich atmosphere
2	1717	silhcoeff	it was sorted under stable flow conditions within a desert
2	1646	silhcoeff	transported by water then weathered in an oxygen rich atmosphere
2	1808	silhcoeff	this rock originated in a windblown desert dune.
2	1833	silhcoeff	the grains are formed from fragments of a pre-existing rock
2	1811	silhcoeff	the it probably originated from the top of a mountain and was carried by a river
2	1812	silhcoeff	it probably originated from the top of a mountain and was carried by a river
2	1175	silhcoeff	this comes from a rock in desert and the grains are oxidised.
2	821	silhcoeff	the sediments are from the desert and they were transported and deposited by wind
2	1892	silhcoeff	formed by weathering caused by the wind in a desert
2	491	silhcoeff	this rock was weathered in arid desert conditions
2	922	silhcoeff	the rock has been transported by air and originated in a desert
2	1131	silhcoeff	the rock originated in an arid desert climate
2	595	silhcoeff	it was formed in oxidising conditions- possibly in a desert environment
2	1334	silhcoeff	chemical weathering in hot humid africa
2	1102	silhcoeff	transported by water. oxidation of minerals.
2	1285	silhcoeff	the sandstone would have been formed in desert conditions.
2	316	silhcoeff	the sandstone was formed from an oxidizing environment in a desert.
2	683	silhcoeff	the origin of the sandstone would be from an arid desert conditon
2	1188	silhcoeff	the rock must have been whethered in a oxygen rich environment.
2	797	silhcoeff	the sandstone originated in desert conditions.
2	1253	silhcoeff	the sandstone originated in desert conditions.
2	1483	silhcoeff	the sandstone originated from desert conditions.
3	603	maxsim	the rounded stones indicate much erosion the redenned grains were formed at high temperatures and high pressure.
3	1675	maxsim	the sandstone would be from a wind blown desert deposit- the red is due to fine grain iron sulfide.
3	603	minsim	the rounded stones indicate much erosion the redenned grains were formed at high temperatures and high pressure.
3	604	minsim	the rounded stones indicate much erosion the redenned grains were formed at high temperatures and high pressure.
3	379	minsize	it was formed in a slow moving water system- in oxygen-rich conditions.
3	1675	maxsize	the sandstone would be from a wind blown desert deposit- the red is due to fine grain iron sulfide.
3	451	silhcoeff	this tells us that the rock was windblown- and probably originated from a sand dune in a desert.

Continua na próxima página

Cluster	Id	Método	Resposta
3	122	silhcoeff	it originated after 2400ma- when atmospheric conditions were oxidising therefore giving the red coloured grains.
3	916	silhcoeff	this sandstone formed in arid and highly oxidative desert conditions- where wind blew the quartz grains.
3	288	silhcoeff	this rock has been a long time having much physical and chemical weathering- pitted by being transported by wind.
3	120	silhcoeff	it originated after 2400ma- when atmospheric conditions were oxidising- therefore giving an abundance of oxidised fe ions.
3	478	silhcoeff	transported by wind- probably from desert sand; contains hematite to give red colour.
3	802	silhcoeff	a high-energy- oxidising environment probably a desert was a feature in this sands past.
4	589	maxsim	formed under arid desert conditions. wind fine sand grains transported by wind. oxidising conditions- example of red-beds
4	1304	maxsim	desert sand- transported by wind- bounced on ground to create pitted surface which sorts them- are compressed into sandstone.
4	86	minsim	high energy transportation and deposition in oxidising conditions on land formed a red-bed- so is probably desert sandstone.
4	87	minsim	high energy transportation by wind- deposition in oxidising conditions on land forming a red-bed- so probably desert sandstone.
4	86	minsize	high energy transportation and deposition in oxidising conditions on land formed a red-bed- so is probably desert sandstone.
4	589	maxsize	formed under arid desert conditions. wind fine sand grains transported by wind. oxidising conditions- example of red-beds
5	1596	maxsim	involved in low tectonic activity- transported in water- rounded from erosion- red from association with iron oxides and weathering
5	310	maxsim	they were formed by desert conditions- wind repeated impacts caused a frosted appearance. there are iron minerals present.
5	371	minsim	origins are arid desert conditions with rounded pitted grains caused by wind transportation and the red surfaces by iron oxide.
5	1608	minsim	red haematite coating/cement from arid environment- probably a desert- and grains rounded and pitted by wind transportation
5	326	minsize	it has come from arid desert- the red colour shows oxidation and pitted grains shows the rock has been frosted.
5	269	maxsize	it originates from arid desert conditions- well -round indicates ancient sandstone- reddened from coating of iron oxide from oxidising conditions.

Continua na próxima página

Cluster	Id	Método	Resposta
5	1510	silhcoeff	found in arid desert conditions - result of wind transportation. iron oxide coating due to oxidising conditions.
			... Foram omitidos os demais resultados (58 *clusters* no total).

Última página

Tabela 15 – Seleção de amostras aplicada para a atividade *Sandstone*.

B.4 Características

A frequência é algo que caracteriza a tendência das respostas ou detalhes de seu conteúdo. Por conta disso, podemos identificar características que foram fundamentais para cada uma das notas atribuídas. Nesse aspecto, a Tabela 16 apresenta os dois extremos de frequência das características das respostas da atividade exemplo *Sandstone.*.

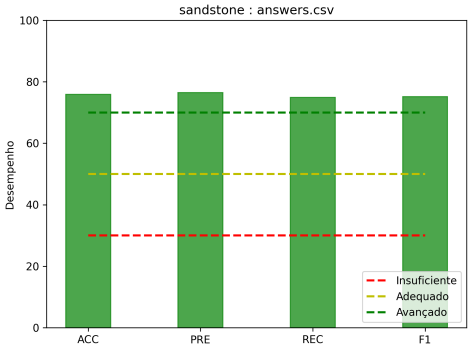
Palavra	Frequência
number	27585
sing	26945
prop	18068
nountype	18068
noun	9400
degree	1880
pos	1855
person	1798
verbform	1749
adj	1673
propn	1662
punct	1519
punctttype	1501
org	1402
verb	1350
tense	1334
peri	1219
desert	1188
fin	879
wind	841
past	806
plur	640
part	614
aspect	599
condit	575

Palavra	Frequência
arid-desert	1
area-wind	1
appearance-	1
ancientdesert	1
ancient-	1
america	1
alon	1
almost	1
airborn	1
aid	1
agent	1
africa	1
affect	1
aeolian	1
actual	1
activity-	1
accuml	1
account	1
absenc	1
abscenc	1
abrass	1
abrad	1
25-	1
2400ma	1
2400	1

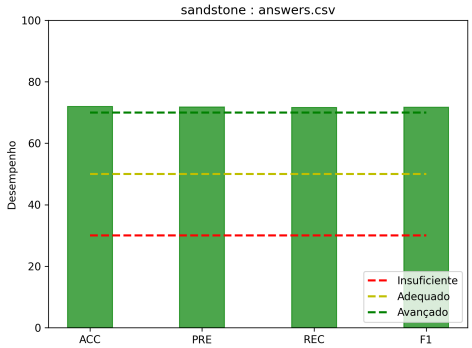
Tabela 16 – Características mais frequentes e menos frequentes encontradas nas respostas da atividade.

B.5 Classificação

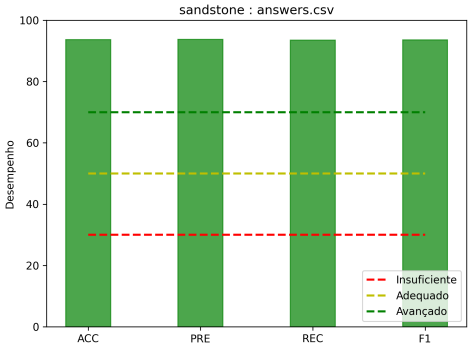
Com as amostras avaliadas, os 6 classificadores são testados em todos os casos, e selecionados conforme sua proximidade no método avaliativo com o conjunto conhecido. Os resultados de classificação e a respectiva matriz de confusão para cada um dos classificadores é apresentada nas Figuras 15 e 16.



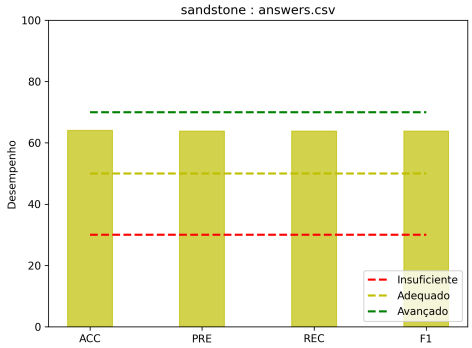
(a) SVM



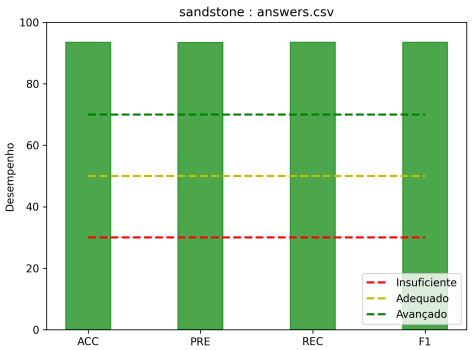
(b) KNN



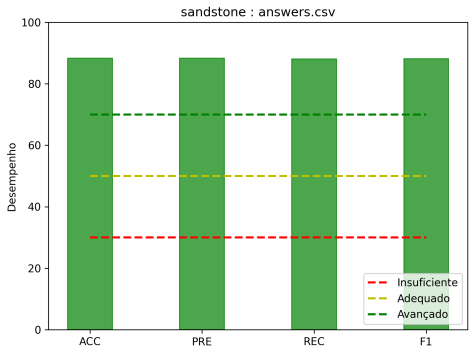
(c) DTR



(d) WSD

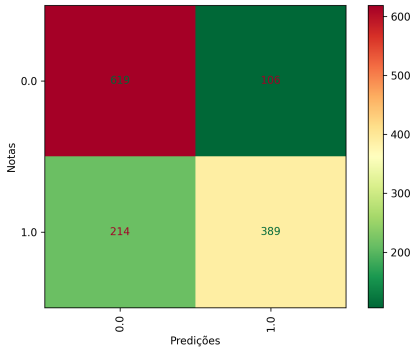


(e) GBC

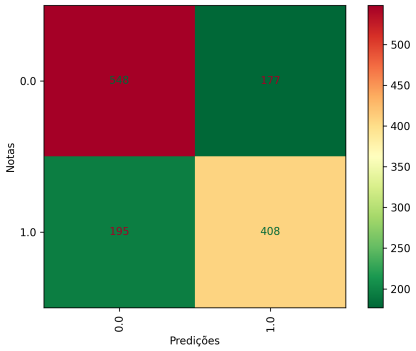


(f) RDF

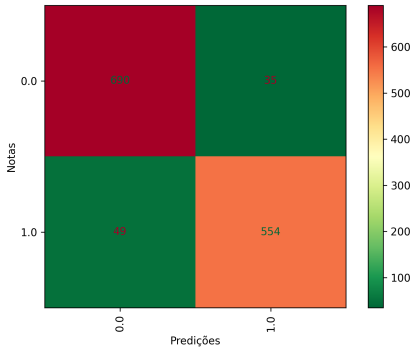
Figura 15 – Resultados de todos os 6 algoritmos de classificação para a atividade exemplo.



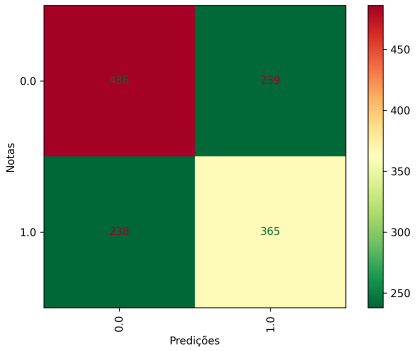
(a) SVM



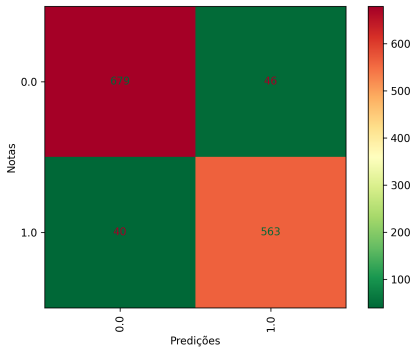
(b) KNN



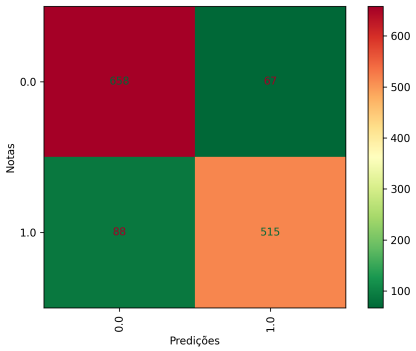
(c) DTR



(d) WSD



(e) GBC



(f) RDF

Figura 16 – Matriz de confusão de todos os 6 algoritmos de classificação para a atividade exemplo.

APÊNDICE C – Experimentos

Os conjuntos de dados são representados pela média de todos os resultados observandos em sua série de atividades. Para ilustrar os resultados de toda série de experimentos foram selecionados a partição de treino e teste com 75% do conjunto de respostas. Nesse aspecto, apresentamos neste capítulo todos os resultados individuais para as atividades de cada base de dados.

C.1 UNT

Atividades da *University of North Texas*

University of North Texas									
					Métricas				
Método	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
					1.1				
	A1			A2			Média		
LINR	0,4099	0,3663	0,6052	0,6200	0,5400	0,7400	0,4025	0,3372	0,5807
LSSR	1,1131	1,4708	1,2128	1,3800	2,1000	1,4500	1,1488	1,6430	1,2818
KNRG	0,3750	0,4028	0,6346	0,6700	0,5600	0,7500	0,3958	0,2326	0,4823
DTRG	0,2500	0,2500	0,5000	0,7500	1,7500	1,3200	0,6875	1,4063	1,1859
WSRG	0,9922	1,1155	1,0562	1,2000	1,5700	1,2500	0,9792	1,1898	1,0908
					1.2				
	A1			A2			Média		
LINR	1,0123	1,3455	1,1599	1,2000	1,9300	1,3900	1,1037	1,5118	1,2295
LSSR	0,9881	1,4546	1,2061	1,3100	2,1800	1,4800	1,1488	1,6430	1,2818
KNRG	1,0833	1,5833	1,2583	1,0800	1,9200	1,3800	1,0833	1,5833	1,2583
DTRG	0,6250	0,8750	0,9354	1,6300	4,8800	2,2100	1,0625	1,5313	1,2374
WSRG	0,9784	1,4262	1,1943	1,3300	2,2900	1,5100	1,1581	1,6863	1,2986
					1.3				
	A1			A2			Média		
LINR	0,1817	0,0865	0,2942	0,2000	0,0400	0,2100	0,1532	0,0519	0,2279
LSSR	0,7440	0,6630	0,8142	0,1300	0,1300	0,3500	0,4345	0,2982	0,5461
KNRG	0,7500	0,6111	0,7817	0,5800	0,3900	0,6200	0,6667	0,4444	0,6667
DTRG	0,1250	0,1250	0,3536	0,6300	0,6300	0,7900	0,8750	1,0000	1,0000
WSRG	0,4734	0,3298	0,5743	0,1800	0,0900	0,3000	0,2456	0,1522	0,3901
					1.4				

Continua na próxima página

University of North Texas

	Métricas								
Método	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
	A1			A2			Média		
LINR	0,6093	0,4189	0,6472	0,2300	0,0600	0,2400	0,3911	0,1654	0,4067
LSSR	0,4762	0,2268	0,4762	0,2400	0,0600	0,2400	0,3571	0,1276	0,3571
KNRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
DTRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	1,8750	6,2500	2,5000
WSRG	0,5916	0,3528	0,5939	0,3300	0,1100	0,3300	0,4709	0,2297	0,4792

1.5

	A1			A2			Média		
LINR	0,7835	1,0148	1,0074	0,1700	0,0500	0,2200	0,3842	0,2369	0,4867
LSSR	1,1488	1,4130	1,1887	0,1900	0,0400	0,1900	0,6220	0,4026	0,6345
KNRG	0,6667	1,4722	1,2134	0,0000	0,0000	0,0000	0,3333	0,3681	0,6067
DTRG	0,8750	2,1250	1,4577	0,0000	0,0000	0,0000	0,3125	0,2813	0,5303
WSRG	1,1383	1,3889	1,1785	0,2600	0,0700	0,2700	0,6260	0,4227	0,6502

1.6

	A1			A2			Média		
LINR	1,2664	2,1394	1,4627	1,0600	1,6200	1,2700	1,1480	1,6028	1,2660
LSSR	1,3750	2,6250	1,6202	1,0700	1,9000	1,3800	1,2232	1,9859	1,4092
KNRG	1,2917	2,7361	1,6541	1,0000	1,3900	1,1800	1,1458	1,7465	1,3216
DTRG	0,7500	0,7500	0,8660	0,7500	1,0000	1,0000	0,6875	0,7188	0,8478
WSRG	1,2882	2,1746	1,4747	0,9600	1,5400	1,2400	1,0739	1,5681	1,2522

1.7

	A1			A2			Média		
LINR	0,4347	0,4249	0,6518	0,2200	0,1000	0,3200	0,2860	0,2110	0,4593
LSSR	0,5893	0,4872	0,6980	0,3700	0,1500	0,3900	0,4405	0,2671	0,5169
KNRG	0,3750	0,6250	0,7906	0,1300	0,1300	0,3500	0,2500	0,3125	0,5590
DTRG	0,3750	0,6250	0,7906	0,1300	0,1300	0,3500	0,2500	0,3125	0,5590
WSRG	0,6966	0,5516	0,7427	0,5100	0,2600	0,5100	0,5573	0,3490	0,5908

10.1

	A1			A2			Média		
LINR	0,5638	0,5504	0,7419	0,2200	0,1200	0,3500	0,3336	0,1746	0,4179
LSSR	0,8148	1,0494	1,0244	0,2400	0,1400	0,3800	0,4537	0,3873	0,6224
KNRG	1,0000	1,1481	1,0715	0,1700	0,1700	0,4100	0,5833	0,4120	0,6419
DTRG	1,1667	2,1667	1,4720	0,1700	0,1700	0,4100	0,9167	1,2083	1,0992
WSRG	0,6499	0,8560	0,9252	0,2500	0,1200	0,3500	0,4184	0,2914	0,5398

10.2

Continua na próxima página

University of North Texas

	Métricas								
Método	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
	A1			A2			Média		
LINR	0,5943	0,3697	0,6080	0,2400	0,0600	0,2500	0,4198	0,1840	0,4290
LSSR	1,3333	1,7778	1,3333	0,3300	0,1100	0,3300	0,8333	0,6944	0,8333
KNRG	0,6111	0,4630	0,6804	0,1700	0,0600	0,2400	0,3889	0,1944	0,4410
DTRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	1,2500	1,8750	1,3693
WSRG	1,0883	1,1853	1,0887	0,2500	0,0600	0,2500	0,6771	0,4594	0,6778

10.3

	A1			A2			Média		
LINR	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
LSSR	0,8889	0,7901	0,8889	0,1100	0,0100	0,1100	0,5000	0,2500	0,5000
KNRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
DTRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
WSRG	0,3419	0,1169	0,3419	0,0500	0,0000	0,0500	0,1955	0,0384	0,1961

10.4

	A1			A2			Média		
LINR	0,4776	0,2529	0,5029	0,0900	0,0100	0,1100	0,2864	0,0925	0,3042
LSSR	0,5556	0,3086	0,5556	0,1100	0,0100	0,1100	0,3333	0,1111	0,3333
KNRG	0,3333	0,1111	0,3333	0,0000	0,0000	0,0000	0,1667	0,0278	0,1667
DTRG	0,5000	0,5000	0,7071	0,0000	0,0000	0,0000	0,0833	0,0417	0,2041
WSRG	0,3508	0,1265	0,3556	0,0700	0,0100	0,0800	0,2188	0,0489	0,2211

10.5

	A1			A2			Média		
LINR	0,9089	0,8620	0,9284	0,6500	0,5100	0,7100	0,7825	0,6522	0,8076
LSSR	1,1667	1,5000	1,2247	0,7000	0,6000	0,7800	0,9352	0,9985	0,9992
KNRG	0,3333	0,1111	0,3333	0,5000	0,3900	0,6200	0,3611	0,1528	0,3909
DTRG	1,1667	4,1667	2,0412	0,8300	0,8300	0,9100	0,8333	2,0833	1,4434
WSRG	1,2418	1,5750	1,2550	0,7700	0,6400	0,8000	1,0103	1,0661	1,0325

10.6

	A1			A2			Média		
LINR	1,0346	1,2264	1,1074	0,4400	0,2300	0,4800	0,6623	0,5535	0,7439
LSSR	1,2037	1,7809	1,3345	0,6100	0,3900	0,6200	0,8426	0,8897	0,9432
KNRG	1,0556	1,8333	1,3540	0,3900	0,3900	0,6200	0,7222	0,8796	0,9379
DTRG	1,5000	5,5000	2,3452	0,1700	0,1700	0,4100	1,0000	2,5833	1,6073
WSRG	1,2254	2,0062	1,4164	0,6700	0,4900	0,7000	0,9010	1,0558	1,0275

10.7

	A1			A2			Média		

Continua na próxima página

University of North Texas

	Métricas								
Método	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
LINR	2,2939	6,9302	2,6325	0,5100	0,4000	0,6300	1,3697	2,2218	1,4906
LSSR	2,2037	5,5586	2,3577	0,6100	0,3900	0,6200	1,2963	1,7508	1,3232
KNRG	2,1111	7,7778	2,7889	0,4400	0,5600	0,7500	1,2222	2,6204	1,6188
DTRG	2,8333	13,1667	3,6286	0,6700	1,0000	1,0000	1,9167	6,5417	2,5577
WSRG	2,1879	5,6644	2,3800	0,6700	0,5000	0,7100	1,3286	1,8067	1,3441

11.1

	A1			A2			Média		
LINR	2,1158	6,7044	2,5893	1,6100	3,2400	1,8000	0,7897	0,7621	0,8730
LSSR	2,6705	8,8626	2,9770	2,1500	5,3800	2,3200	1,0284	1,1663	1,0800
KNRG	1,6667	6,7778	2,6034	1,4000	3,0000	1,7300	0,6667	0,8403	0,9167
DTRG	0,8750	2,1250	1,4577	0,8100	1,8400	1,3600	0,3750	0,3750	0,6124
WSRG	2,5181	8,8258	2,9708	2,0900	5,0600	2,2500	0,9877	1,0906	1,0443

11.10

	A1			A2			Média		
LINR	1,0057	1,3266	1,1518	1,0100	1,3300	1,1500	0,2211	0,0631	0,2512
LSSR	0,8182	0,6694	0,8182	0,8200	0,6700	0,8200	0,1818	0,0331	0,1818
KNRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
DTRG	0,7500	4,5000	2,1213	0,8800	6,1300	2,4700	0,1875	0,2813	0,5303
WSRG	1,1894	1,4233	1,1930	1,1700	1,3800	1,1700	0,2603	0,0682	0,2612

11.2

	A1			A2			Média		
LINR	0,4458	0,2450	0,4949	0,4300	0,2100	0,4500	0,2186	0,0611	0,2472
LSSR	1,3636	1,8595	1,3636	1,0300	1,0800	1,0400	0,5455	0,2975	0,5455
KNRG	0,0000	0,0000	0,0000	0,1300	0,0600	0,2500	0,0000	0,0000	0,0000
DTRG	0,0000	0,0000	0,0000	0,1300	0,0600	0,2500	0,0000	0,0000	0,0000
WSRG	1,3046	1,7060	1,3061	1,0200	1,0600	1,0300	0,5517	0,3045	0,5519

11.3

	A1			A2			Média		
LINR	3,1541	13,2962	3,6464	1,4400	4,1000	2,0300	0,7615	1,0296	1,0147
LSSR	2,5682	8,4174	2,9013	1,8600	4,9400	2,2200	0,6761	0,6994	0,8363
KNRG	4,0000	19,5556	4,4222	1,6500	6,8400	2,6200	1,1042	2,0590	1,4349
DTRG	2,5000	18,0000	4,2426	2,9400	13,4100	3,6600	1,4375	2,3438	1,5309
WSRG	2,1943	6,4104	2,5319	2,1200	5,6300	2,3700	0,5240	0,5377	0,7332

11.4

	A1			A2			Média		
LINR	2,5337	8,9407	2,9901	0,7500	0,6500	0,8100	0,5930	0,5115	0,7152

Continua na próxima página

University of North Texas									
					Métricas				
Método	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
LSSR	2,4886	8,4948	2,9146	0,7800	0,6400	0,8000	0,5852	0,4765	0,6903
KNRG	2,4583	11,0139	3,3187	0,7500	0,9200	0,9600	0,6458	0,7674	0,8760
DTRG	1,8750	11,3750	3,3727	1,6300	4,5600	2,1400	1,0000	1,6875	1,2990
WSRG	2,5430	8,5286	2,9204	0,8800	0,7900	0,8900	0,6335	0,4987	0,7062

11.5

	A1			A2			Média		
LINR	0,8766	1,0154	1,0077	0,5400	0,4600	0,6800	0,3160	0,1170	0,3420
LSSR	0,9318	0,8719	0,9338	0,5500	0,3000	0,5500	0,3182	0,1049	0,3238
KNRG	0,4583	0,8472	0,9204	0,0000	0,0000	0,0000	0,1042	0,0451	0,2125
DTRG	1,0000	8,0000	2,8284	0,0000	0,0000	0,0000	0,5625	2,5313	1,5910
WSRG	1,1376	1,3610	1,1666	0,6700	0,4600	0,6800	0,4205	0,1867	0,4321

11.6

	A1			A2			Média		
LINR	0,2253	0,1158	0,3403	0,5300	0,4600	0,6800	0,1584	0,0426	0,2065
LSSR	0,2273	0,0517	0,2273	0,6400	0,4000	0,6400	0,1818	0,0331	0,1818
KNRG	0,2083	0,3472	0,5893	0,2100	0,3500	0,5900	0,0833	0,0556	0,2357
DTRG	0,0000	0,0000	0,0000	0,6300	3,1300	1,7700	0,1250	0,1250	0,3536
WSRG	0,2004	0,0405	0,2011	0,4200	0,1800	0,4200	0,1267	0,0161	0,1270

11.7

	A1			A2			Média		
LINR	3,4045	14,8073	3,8480	0,3200	0,2000	0,4500	0,8401	0,8683	0,9318
LSSR	4,2273	20,9050	4,5722	0,2000	0,2700	0,5200	1,0114	1,1524	1,0735
KNRG	3,5833	15,7778	3,9721	0,6900	0,8000	0,8900	0,9583	1,0347	1,0172
DTRG	4,1250	31,3750	5,6013	0,9400	1,7800	1,3300	0,5000	0,5000	0,7071
WSRG	3,6914	16,4666	4,0579	0,3700	0,3500	0,6000	0,8440	0,8739	0,9348

11.8

	A1			A2			Média		
LINR	3,4914	15,0517	3,8796	1,3000	1,9400	1,3900	1,0929	1,5992	1,2646
LSSR	3,7614	18,3130	4,2794	1,5500	2,5600	1,6000	1,3580	2,0566	1,4341
KNRG	3,8333	25,3333	5,0332	1,5800	3,1700	1,7800	1,2917	2,7014	1,6436
DTRG	1,7500	7,5000	2,7386	1,5600	3,3400	1,8300	0,5625	0,6563	0,8101
WSRG	3,5744	16,7159	4,0885	1,4400	2,3000	1,5200	1,2580	1,8765	1,3699

11.9

	A1			A2			Média		
LINR	0,5870	0,4736	0,6882	0,1600	0,0300	0,1800	0,2180	0,0783	0,2798
LSSR	0,8068	0,7603	0,8720	0,1600	0,0300	0,1600	0,2898	0,1113	0,3336

Continua na próxima página

University of North Texas

	Métricas								
Método	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
KNRG	0,6250	0,6250	0,7906	0,2900	0,0900	0,3000	0,1250	0,0347	0,1863
DTRG	2,5000	9,7500	3,1225	0,6300	0,6300	0,7900	0,5625	0,4063	0,6374
WSRG	0,9529	1,0122	1,0061	0,0700	0,0100	0,0700	0,3265	0,1378	0,3713

12.1

	A1			A2			Média		
LINR	1,3203	5,6592	2,3789	0,0400	0,0000	0,0600	0,3252	0,3738	0,6114
LSSR	1,5102	8,2132	2,8659	0,0700	0,0100	0,0700	0,3707	0,5176	0,7194
KNRG	0,9048	4,6190	2,1492	0,0700	0,0400	0,1900	0,2619	0,3373	0,5808
DTRG	1,4286	9,4286	3,0706	0,0000	0,0000	0,0000	0,0714	0,0357	0,1890
WSRG	1,4695	7,5071	2,7399	0,0900	0,0100	0,0900	0,3847	0,4869	0,6978

12.10

	A1			A2			Média		
LINR	3,8698	16,6488	4,0803	1,2500	1,7600	1,3300	1,2352	1,7781	1,3334
LSSR	3,5442	14,5737	3,8176	1,3600	1,8400	1,3600	1,2857	1,7755	1,3325
KNRG	4,2857	20,9524	4,5774	1,3800	2,3600	1,5400	1,4286	2,2619	1,5040
DTRG	2,1429	17,8571	4,2258	1,1400	4,5700	2,1400	1,1429	3,6429	1,9086
WSRG	3,5405	14,2504	3,7750	1,2900	1,6800	1,3000	1,2469	1,6961	1,3024

12.2

	A1			A2			Média		
LINR	2,8727	9,7189	3,1175	1,0200	1,6500	1,2800	0,8079	0,8298	0,9109
LSSR	2,7415	8,4512	2,9071	1,0700	2,0400	1,4300	0,7075	0,7336	0,8565
KNRG	2,2857	10,0000	3,1623	1,1700	1,7300	1,3100	0,6429	0,6944	0,8333
DTRG	2,2857	11,7143	3,4226	1,0000	1,7100	1,3100	1,5000	2,4643	1,5698
WSRG	2,8682	10,1307	3,1829	1,6500	3,6300	1,9000	0,8761	1,0963	1,0471

12.3

	A1			A2			Média		
LINR	2,3451	6,6527	2,5793	0,4400	0,3300	0,5700	0,6863	0,6035	0,7769
LSSR	3,1905	10,9955	3,3159	1,1600	1,6900	1,3000	1,0536	1,2905	1,1360
KNRG	2,5714	7,4286	2,7255	0,9000	0,8500	0,9200	0,8690	0,8209	0,9061
DTRG	2,4286	11,0000	3,3166	1,3600	2,1800	1,4800	1,4107	2,0558	1,4338
WSRG	3,4947	12,3186	3,5098	1,0000	1,2300	1,1100	1,2095	1,5382	1,2402

12.4

	A1			A2			Média		
LINR	3,3970	12,3565	3,5152	0,8000	0,8300	0,9100	0,9074	0,9882	0,9941
LSSR	3,7143	13,7959	3,7143	0,9100	1,1700	1,0800	1,0238	1,1094	1,0533
KNRG	1,1429	3,0476	1,7457	0,9800	1,1100	1,0500	0,4524	0,2976	0,5455

Continua na próxima página

University of North Texas									
					Métricas				
Método	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
DTRG	1,1429	9,1429	3,0237	1,2900	2,5700	1,6000	0,0714	0,0357	0,1890
WSRG	3,2987	10,8947	3,3007	0,8800	1,1400	1,0700	0,9084	0,8895	0,9431

12.5

	A1			A2			Média		
LINR	0,1996	0,0451	0,2125	0,2000	0,0500	0,2100	0,0408	0,0021	0,0453
LSSR	0,1429	0,0204	0,1429	0,1400	0,0200	0,1400	0,0238	0,0006	0,0238
KNRG	0,3333	0,2063	0,4543	0,3300	0,2100	0,4500	0,0714	0,0119	0,1091
DTRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
WSRG	0,0939	0,0089	0,0941	0,0900	0,0100	0,0900	0,0158	0,0003	0,0158

12.6

	A1			A2			Média		
LINR	0,3505	0,1884	0,4341	0,7800	3,4900	1,8700	0,2316	0,1547	0,3933
LSSR	1,3810	1,9070	1,3810	0,8800	3,2900	1,8100	0,4150	0,1791	0,4232
KNRG	0,0000	0,0000	0,0000	0,7100	3,5700	1,8900	0,1429	0,1429	0,3780
DTRG	0,0000	0,0000	0,0000	0,7100	3,5700	1,8900	0,1429	0,1429	0,3780
WSRG	1,6368	2,6858	1,6388	0,9300	3,2500	1,8000	0,4715	0,2248	0,4742

12.7

	A1			A2			Média		
LINR	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
LSSR	0,1905	0,0363	0,1905	0,1900	0,0400	0,1900	0,0238	0,0006	0,0238
KNRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
DTRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
WSRG	0,1749	0,0306	0,1749	0,1700	0,0300	0,1700	0,0333	0,0011	0,0333

12.8

	A1			A2			Média		
LINR	2,4115	9,0905	3,0150	1,0700	1,3700	1,1700	0,7580	0,8569	0,9257
LSSR	2,6667	10,3492	3,2170	1,3600	2,3200	1,5200	0,8707	1,1519	1,0733
KNRG	3,1429	14,1270	3,7586	1,3800	2,1900	1,4800	0,9524	1,2222	1,1055
DTRG	2,1429	9,0000	3,0000	0,7100	0,9300	0,9600	0,8571	1,1429	1,0690
WSRG	2,7338	9,8338	3,1359	1,3800	2,3900	1,5500	0,8702	1,1198	1,0582

12.9

	A1			A2			Média		
LINR	3,8243	16,4647	4,0577	0,8300	1,3000	1,1400	1,0749	1,3632	1,1676
LSSR	4,3537	20,4989	4,5276	1,0900	2,1100	1,4500	1,2925	1,8730	1,3686
KNRG	3,2857	16,6825	4,0844	1,2900	2,1700	1,4700	1,0000	1,8175	1,3481
DTRG	3,4286	15,7143	3,9641	0,7900	1,5400	1,2400	1,1429	2,1429	1,4639

Continua na próxima página

University of North Texas									
					Métricas				
Método	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
WSRG	4,4164	22,3348	4,7260	1,4100	2,9600	1,7200	1,3086	2,0614	1,4358

2.1

	A1			A2			Média		
LINR	0,7424	1,1739	1,0835	0,3700	0,3000	0,5500	0,4524	0,5808	0,7621
LSSR	1,2273	1,9959	1,4128	0,6300	0,4900	0,7000	0,8068	0,9788	0,9894
KNRG	0,5000	1,4444	1,2019	0,5000	0,5300	0,7300	0,5000	0,8542	0,9242
DTRG	0,5000	1,2500	1,1180	0,1300	0,1300	0,3500	0,2500	0,5000	0,7071
WSRG	1,1674	1,8467	1,3589	0,5700	0,4200	0,6500	0,7506	0,8430	0,9182

2.2

	A1			A2			Média		
LINR	1,2575	2,4373	1,5612	0,3800	0,2500	0,5000	0,5849	0,6973	0,8351
LSSR	1,3295	2,7831	1,6682	0,4100	0,2900	0,5400	0,6193	0,8215	0,9064
KNRG	1,5000	3,3889	1,8409	0,4200	0,2800	0,5300	0,7917	1,0139	1,0069
DTRG	1,6250	4,3750	2,0917	0,2500	0,2500	0,5000	1,0625	2,0313	1,4252
WSRG	1,2952	2,7708	1,6646	0,4100	0,2900	0,5400	0,6031	0,8114	0,9008

2.3

	A1			A2			Média		
LINR	0,7580	0,7313	0,8552	0,6900	0,7300	0,8500	0,5122	0,3437	0,5863
LSSR	0,9659	1,2004	1,0956	0,7700	0,8500	0,9200	0,5511	0,4680	0,6841
KNRG	0,7917	1,0139	1,0069	0,6300	0,7100	0,8400	0,7083	0,6736	0,8207
DTRG	1,3750	2,1250	1,4577	0,3800	0,3800	0,6100	0,6250	0,6250	0,7906
WSRG	0,9988	1,1968	1,0940	0,7000	0,7500	0,8700	0,5225	0,4005	0,6329

2.4

	A1			A2			Média		
LINR	0,4294	0,2428	0,4927	0,1800	0,0500	0,2300	0,3086	0,1334	0,3652
LSSR	0,7727	0,5971	0,7727	0,3600	0,1300	0,3600	0,5682	0,3228	0,5682
KNRG	0,0833	0,0556	0,2357	0,0800	0,0600	0,2400	0,0833	0,0556	0,2357
DTRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
WSRG	0,8200	0,6794	0,8243	0,3300	0,1100	0,3300	0,5508	0,3100	0,5568

2.5

	A1			A2			Média		
LINR	1,3432	2,5149	1,5859	0,8600	1,1800	1,0900	1,0639	1,6266	1,2754
LSSR	1,9318	3,7459	1,9354	1,5100	2,4000	1,5500	1,6364	2,7960	1,6721
KNRG	1,0000	2,0556	1,4337	0,5000	0,8100	0,9000	0,7500	1,2431	1,1149
DTRG	1,0000	4,0000	2,0000	1,0000	2,5000	1,5000	0,7500	2,3125	1,5207
WSRG	2,1447	4,7530	2,1801	1,9400	4,1300	2,0300	1,8760	4,1829	2,0452

Continua na próxima página

University of North Texas

	Métricas								
Método	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
				2.6					
	A1			A2			Média		
LINR	0,5807	0,3864	0,6216	0,3200	0,1600	0,4000	0,3454	0,1568	0,3959
LSSR	0,6250	0,6250	0,7906	0,3900	0,3400	0,5900	0,4886	0,3528	0,5940
KNRG	0,5833	0,6111	0,7817	0,3300	0,1900	0,4400	0,3750	0,2431	0,4930
DTRG	1,1250	3,3750	1,8371	0,2500	0,2500	0,5000	0,3750	0,2500	0,5000
WSRG	0,5871	0,4694	0,6851	0,3700	0,3400	0,5900	0,4607	0,3110	0,5576
				2.7					
	A1			A2			Média		
LINR	1,3229	2,4271	1,5579	0,5200	0,3400	0,5800	0,7363	0,8278	0,9098
LSSR	1,5000	3,0000	1,7321	0,4100	0,2900	0,5400	0,8864	1,0217	1,0108
KNRG	0,8750	1,5139	1,2304	0,6700	0,5600	0,7500	0,6458	0,6563	0,8101
DTRG	2,5000	8,0000	2,8284	1,5000	2,7500	1,6600	1,4375	3,2188	1,7941
WSRG	1,5978	3,2000	1,7888	0,3900	0,3500	0,5900	0,9730	1,1423	1,0688
				3.1					
	A1			A2			Média		
LINR	0,4077	0,2882	0,5369	0,0200	0,0000	0,0400	0,2016	0,0747	0,2733
LSSR	1,1739	1,3781	1,1739	0,0900	0,0100	0,0900	0,6304	0,3974	0,6304
KNRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
DTRG	0,1250	0,1250	0,3536	0,0000	0,0000	0,0000	0,0625	0,0313	0,1768
WSRG	0,7787	0,6222	0,7888	0,0400	0,0000	0,0400	0,3927	0,1584	0,3981
				3.2					
	A1			A2			Média		
LINR	0,9677	1,4515	1,2048	1,1900	1,7600	1,3300	0,4730	0,3416	0,5845
LSSR	1,0707	1,2691	1,1266	1,2600	1,5900	1,2600	0,5353	0,3173	0,5633
KNRG	0,9583	2,0972	1,4482	1,3300	2,7200	1,6500	0,4792	0,5243	0,7241
DTRG	1,6250	7,3750	2,7157	2,0000	9,5000	3,0800	0,8125	1,8438	1,3578
WSRG	0,9202	0,9887	0,9943	1,0500	1,1300	1,0600	0,4631	0,2510	0,5010
				3.3					
	A1			A2			Média		
LINR	1,2649	2,8550	1,6897	0,9100	1,4700	1,2100	0,9681	1,5251	1,2350
LSSR	1,7935	4,4750	2,1154	1,1700	2,1300	1,4600	1,4130	2,7395	1,6551
KNRG	1,1667	2,2222	1,4907	1,1700	2,3100	1,5200	1,0000	1,7361	1,3176
DTRG	1,5000	4,7500	2,1794	1,0000	1,5000	1,2200	1,3125	3,2813	1,8114
WSRG	1,5709	3,3694	1,8356	0,9500	1,5700	1,2500	1,2198	1,9614	1,4005

Continua na próxima página

University of North Texas

	Métricas								
Método	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
				3.4					
	A1			A2			Média		
LINR	0,6930	0,5564	0,7459	0,6400	0,5400	0,7300	0,6643	0,5166	0,7187
LSSR	1,0000	1,0000	1,0000	0,7500	0,6900	0,8300	0,8750	0,7983	0,8935
KNRG	1,0417	1,2639	1,1242	0,8800	1,1000	1,0500	0,9583	1,1597	1,0769
DTRG	0,3750	0,6250	0,7906	0,6300	0,8800	0,9400	0,0625	0,0313	0,1768
WSRG	0,9493	0,9099	0,9539	0,7300	0,6600	0,8100	0,8418	0,7539	0,8683
				3.5					
	A1			A2			Média		
LINR	0,4770	0,4571	0,6761	0,1900	0,0500	0,2200	0,3345	0,1961	0,4428
LSSR	1,2609	1,5898	1,2609	0,5200	0,2700	0,5200	0,8913	0,7944	0,8913
KNRG	0,1667	0,1111	0,3333	0,0000	0,0000	0,0000	0,0833	0,0278	0,1667
DTRG	0,7500	2,2500	1,5000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
WSRG	1,0349	1,1112	1,0541	0,5100	0,2600	0,5100	0,7263	0,5507	0,7421
				3.6					
	A1			A2			Média		
LINR	0,4473	0,2482	0,4982	0,0100	0,0000	0,0200	0,2289	0,0654	0,2557
LSSR	0,9130	0,8336	0,9130	0,0900	0,0100	0,0900	0,5000	0,2500	0,5000
KNRG	0,1667	0,1389	0,3727	0,0000	0,0000	0,0000	0,0833	0,0347	0,1863
DTRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0625	0,0313	0,1768
WSRG	0,7443	0,5604	0,7486	0,0800	0,0100	0,0800	0,4006	0,1630	0,4037
				3.7					
	A1			A2			Média		
LINR	0,8118	0,9148	0,9565	0,3500	0,2400	0,4900	0,5764	0,4483	0,6696
LSSR	1,3587	1,9060	1,3806	0,5300	0,2900	0,5400	0,9457	0,9031	0,9503
KNRG	0,7500	2,2500	1,5000	0,2500	0,2500	0,5000	0,5000	1,0000	1,0000
DTRG	0,7500	2,2500	1,5000	0,5000	0,7500	0,8700	1,0000	2,5000	1,5811
WSRG	1,2888	1,6982	1,3032	0,5500	0,3700	0,6100	0,9054	0,8331	0,9127
				4.2					
	A1			A2			Média		
LINR	1,2413	1,9135	1,3833	0,3100	0,1200	0,3500	0,6552	0,5730	0,7570
LSSR	1,4091	2,3595	1,5361	0,3600	0,1500	0,3800	0,7500	0,7583	0,8708
KNRG	0,9167	1,2500	1,1180	0,5000	0,3300	0,5800	0,6250	0,5556	0,7454
DTRG	1,5000	3,7500	1,9365	0,1300	0,1300	0,3500	0,8125	1,2188	1,1040
WSRG	1,3864	2,3319	1,5271	0,3200	0,1300	0,3600	0,7401	0,7442	0,8627

Continua na próxima página

University of North Texas

	Métricas								
Método	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
					4.3				
	A1			A2			Média		
LINR	0,1759	0,0854	0,2922	0,1300	0,0500	0,2100	0,1534	0,0645	0,2540
LSSR	0,9545	0,9112	0,9545	0,6800	0,4600	0,6800	0,8182	0,6694	0,8182
KNRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
DTRG	0,2500	0,2500	0,5000	0,0000	0,0000	0,0000	0,5000	1,0000	1,0000
WSRG	0,8381	0,7026	0,8382	0,6400	0,4100	0,6400	0,7360	0,5420	0,7362
					4.4				
	A1			A2			Média		
LINR	1,6441	3,8001	1,9494	0,3500	0,1400	0,3800	0,8285	0,8707	0,9331
LSSR	1,7273	4,4855	2,1179	0,2300	0,0500	0,2300	0,8352	0,9742	0,9870
KNRG	1,8750	5,7361	2,3950	0,0000	0,0000	0,0000	0,9375	1,4340	1,1975
DTRG	1,8750	5,3750	2,3184	0,0000	0,0000	0,0000	0,9375	1,5938	1,2624
WSRG	1,6316	3,4866	1,8672	0,4200	0,1800	0,4200	0,7640	0,7231	0,8504
					4.5				
	A1			A2			Média		
LINR	1,3273	2,1912	1,4803	1,0200	2,6400	1,6300	1,0554	2,1048	1,4508
LSSR	1,9091	3,9959	1,9990	1,1100	3,0500	1,7500	1,3977	3,0253	1,7393
KNRG	0,9583	2,5972	1,6116	1,0000	2,4700	1,5700	0,8958	2,1354	1,4613
DTRG	1,0000	2,7500	1,6583	1,5000	3,2500	1,8000	1,0625	2,2188	1,4895
WSRG	1,7965	3,5847	1,8933	1,1600	2,8200	1,6800	1,3038	2,7296	1,6522
					4.6				
	A1			A2			Média		
LINR	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
LSSR	2,0455	4,1839	2,0455	1,3200	1,7400	1,3200	1,6818	2,8285	1,6818
KNRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
DTRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
WSRG	1,8946	3,5894	1,8946	1,1700	1,3700	1,1700	1,5217	2,3157	1,5217
					4.7				
	A1			A2			Média		
LINR	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
LSSR	1,3182	1,7376	1,3182	0,6400	0,4000	0,6400	0,9773	0,9551	0,9773
KNRG	1,6667	2,7778	1,6667	1,0000	1,0000	1,0000	1,3333	1,7778	1,3333
DTRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
WSRG	1,2431	1,5452	1,2431	0,7200	0,5200	0,7200	0,9848	0,9698	0,9848
					5.1				

Continua na próxima página

University of North Texas

	Métricas								
Método	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
	A1			A2			Média		
LINR	0,8990	1,4030	1,1845	0,5600	0,4600	0,6800	0,5964	0,7381	0,8591
LSSR	1,0680	1,8934	1,3760	0,6500	0,5500	0,7400	0,7551	0,9898	0,9949
KNRG	0,5238	1,3492	1,1616	0,5200	0,4600	0,6800	0,5238	0,7857	0,8864
DTRG	0,5714	2,2857	1,5119	0,5700	0,5700	0,7600	0,5714	1,0000	1,0000
WSRG	1,1017	1,8345	1,3544	0,6300	0,5300	0,7300	0,7297	0,9453	0,9723

5.2

	A1			A2			Média		
LINR	0,3621	0,2379	0,4877	0,3700	0,1600	0,4000	0,1961	0,0420	0,2050
LSSR	0,3878	0,2063	0,4543	0,4500	0,2100	0,4600	0,2415	0,0618	0,2486
KNRG	0,2857	0,2857	0,5345	0,2400	0,1700	0,4200	0,1190	0,0437	0,2089
DTRG	0,2857	0,2857	0,5345	0,7100	0,7100	0,8500	0,1429	0,0714	0,2673
WSRG	0,3949	0,2076	0,4556	0,4300	0,1900	0,4400	0,2354	0,0593	0,2435

5.3

	A1			A2			Média		
LINR	0,9641	1,1114	1,0542	0,3600	0,2200	0,4700	0,4826	0,2606	0,5105
LSSR	1,2449	1,6939	1,3015	0,3300	0,3200	0,5600	0,6259	0,4649	0,6818
KNRG	1,2857	2,0159	1,4198	0,6200	0,4600	0,6800	0,6190	0,4524	0,6726
DTRG	1,2857	4,4286	2,1044	0,7100	0,7100	0,8500	0,7857	0,9643	0,9820
WSRG	1,1513	1,4846	1,2184	0,3700	0,3400	0,5800	0,5842	0,4000	0,6325

5.4

	A1			A2			Média		
LINR	0,4003	0,2045	0,4523	0,1400	0,1400	0,3800	0,1429	0,0714	0,2673
LSSR	1,2381	1,6553	1,2866	0,7600	0,6300	0,8000	0,9762	1,0040	1,0020
KNRG	0,4286	0,5238	0,7237	0,4300	0,2700	0,5200	0,4286	0,3413	0,5842
DTRG	0,1429	0,1429	0,3780	0,5700	1,4300	1,2000	0,4286	0,9286	0,9636
WSRG	1,2276	1,6342	1,2783	0,6700	0,4800	0,6900	0,9322	0,9092	0,9535

6.2

	A1			A2			Média		
LINR	0,0642	0,0177	0,1331	0,3400	0,1700	0,4100	0,1966	0,0541	0,2327
LSSR	0,3684	0,1357	0,3684	0,6300	0,4000	0,6300	0,5000	0,2500	0,5000
KNRG	0,0000	0,0000	0,0000	0,1400	0,0500	0,2200	0,0714	0,0119	0,1091
DTRG	0,0000	0,0000	0,0000	0,4300	0,4300	0,6500	0,0000	0,0000	0,0000
WSRG	0,1592	0,0262	0,1620	0,4700	0,2400	0,4900	0,3294	0,1126	0,3356

6.3

Continua na próxima página

University of North Texas

	Métricas								
Método	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
	A1			A2			Média		
LINR	0,2597	0,1428	0,3779	0,3800	0,2000	0,4500	0,2290	0,0910	0,3016
LSSR	0,4511	0,3182	0,5641	0,4800	0,2500	0,5000	0,3910	0,2008	0,4481
KNRG	0,3810	0,1905	0,4364	0,3800	0,2200	0,4700	0,3333	0,1429	0,3780
DTRG	0,1429	0,1429	0,3780	0,7100	1,0000	1,0000	0,2857	0,2143	0,4629
WSRG	0,3536	0,2256	0,4749	0,3700	0,1900	0,4400	0,3135	0,1156	0,3401

6.4

	A1			A2			Média		
LINR	1,0161	1,9809	1,4075	0,6100	0,4800	0,6900	0,5079	0,3860	0,6213
LSSR	1,1429	1,7143	1,3093	0,8000	0,9300	0,9700	0,5451	0,4475	0,6689
KNRG	1,0476	1,9048	1,3801	0,5700	0,4100	0,6400	0,5238	0,4127	0,6424
DTRG	1,0000	2,7143	1,6475	0,8600	1,4300	1,2000	0,2857	0,2143	0,4629
WSRG	1,1301	1,6561	1,2869	0,7900	1,0500	1,0200	0,5483	0,4646	0,6816

6.5

	A1			A2			Média		
LINR	0,7138	1,4242	1,1934	0,4600	0,6300	0,7900	0,5565	0,5387	0,7340
LSSR	0,8271	1,7159	1,3099	0,8000	1,1000	1,0500	0,6992	0,8876	0,9421
KNRG	0,6667	1,7143	1,3093	0,4800	0,8600	0,9300	0,4762	0,6111	0,7817
DTRG	0,4286	0,7143	0,8452	0,2900	0,2900	0,5300	0,0714	0,0357	0,1890
WSRG	0,7678	1,6226	1,2738	0,7100	1,0000	1,0000	0,6759	0,8143	0,9024

6.6

	A1			A2			Média		
LINR	0,1230	0,0245	0,1565	0,0900	0,0100	0,1200	0,1040	0,0183	0,1354
LSSR	1,9474	3,7922	1,9474	1,4200	2,0200	1,4200	1,6842	2,8366	1,6842
KNRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
DTRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
WSRG	1,7782	3,1751	1,7819	1,2300	1,5100	1,2300	1,5043	2,2684	1,5061

6.7

	A1			A2			Média		
LINR	0,0852	0,0125	0,1118	0,7100	1,7500	1,3200	0,3550	0,3982	0,6310
LSSR	0,1579	0,0249	0,1579	0,7800	1,6600	1,2900	0,4248	0,3764	0,6135
KNRG	0,0952	0,0317	0,1782	0,7100	1,8600	1,3600	0,4048	0,4722	0,6872
DTRG	0,0000	0,0000	0,0000	0,7100	1,8600	1,3600	0,3571	0,4643	0,6814
WSRG	0,1150	0,0136	0,1168	0,7300	1,7700	1,3300	0,3902	0,4069	0,6379

7.1

	A1			A2			Média		

Continua na próxima página

University of North Texas

	Métricas								
Método	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
LINR	0,0848	0,0095	0,0973	0,0900	0,0100	0,1000	0,0797	0,0086	0,0926
LSSR	0,2105	0,0443	0,2105	0,1600	0,0200	0,1600	0,1842	0,0339	0,1842
KNRG	0,2857	0,0952	0,3086	0,0500	0,0200	0,1300	0,1190	0,0198	0,1409
DTRG	0,0000	0,0000	0,0000	0,2900	0,2900	0,5300	0,0714	0,0357	0,1890
WSRG	0,2862	0,0833	0,2887	0,1600	0,0300	0,1600	0,2351	0,0558	0,2361

7.2

	A1			A2			Média		
LINR	0,4224	0,4423	0,6651	0,2500	0,0700	0,2600	0,2996	0,1173	0,3425
LSSR	0,4361	0,4954	0,7039	0,2600	0,0700	0,2600	0,3120	0,1313	0,3623
KNRG	0,2381	0,3968	0,6299	0,0000	0,0000	0,0000	0,1190	0,0992	0,3150
DTRG	0,5714	1,1429	1,0690	0,0000	0,0000	0,0000	0,1429	0,0714	0,2673
WSRG	0,4228	0,4047	0,6361	0,2600	0,0700	0,2600	0,3172	0,1365	0,3694

7.3

	A1			A2			Média		
LINR	1,4713	2,9315	1,7122	0,6300	0,5200	0,7200	0,9552	1,1163	1,0566
LSSR	1,5940	3,2505	1,8029	0,6500	0,5300	0,7300	1,0414	1,2379	1,1126
KNRG	1,7143	4,1587	2,0393	0,7100	0,7100	0,8500	1,1190	1,6944	1,3017
DTRG	2,2857	7,7143	2,7775	0,4300	0,7100	0,8500	0,8571	1,3571	1,1650
WSRG	1,5320	2,7555	1,6600	0,6800	0,5600	0,7500	1,0087	1,0946	1,0462

7.4

	A1			A2			Média		
LINR	0,8589	0,9122	0,9551	0,5700	0,3700	0,6100	0,6270	0,5028	0,7091
LSSR	1,1128	1,4460	1,2025	0,9200	1,0300	1,0200	0,9098	1,0293	1,0145
KNRG	0,8095	1,6349	1,2786	0,9000	0,9700	0,9800	0,8571	1,1111	1,0541
DTRG	0,7143	1,2857	1,1339	0,1400	0,1400	0,3800	0,5714	0,7143	0,8452
WSRG	1,1743	1,5235	1,2343	0,9100	1,1400	1,0700	0,9571	1,1960	1,0936

7.5

	A1			A2			Média		
LINR	0,1572	0,0598	0,2444	0,0800	0,0200	0,1300	0,1190	0,0347	0,1863
LSSR	0,7895	0,6233	0,7895	0,5300	0,2800	0,5300	0,6579	0,4328	0,6579
KNRG	0,0952	0,0635	0,2520	0,1400	0,0800	0,2800	0,1190	0,0675	0,2597
DTRG	0,0000	0,0000	0,0000	0,2900	0,2900	0,5300	0,0000	0,0000	0,0000
WSRG	0,5636	0,3194	0,5651	0,4000	0,1600	0,4000	0,4809	0,2324	0,4820

7.6

	A1			A2			Média		
LINR	0,4784	0,2684	0,5181	0,4500	0,4500	0,6700	0,3166	0,1609	0,4011

Continua na próxima página

University of North Texas									
					Métricas				
Método	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
LSSR	0,4812	0,2319	0,4816	0,3900	0,3600	0,6000	0,2932	0,1336	0,3655
KNRG	0,6190	0,6190	0,7868	0,3800	0,3200	0,5600	0,3095	0,2024	0,4499
DTRG	0,8571	0,8571	0,9258	0,4300	0,7100	0,8500	0,2857	0,1429	0,3780
WSRG	0,5419	0,2979	0,5458	0,3900	0,3600	0,6000	0,2966	0,1261	0,3551

7.7

	A1			A2			Média		
LINR	1,0967	2,6588	1,6306	1,0300	2,7300	1,6500	1,0349	2,6708	1,6343
LSSR	1,4060	3,2580	1,8050	1,2600	3,1000	1,7600	1,3308	3,1486	1,7744
KNRG	0,7619	2,7619	1,6619	0,8100	2,8100	1,6800	0,7381	2,7500	1,6583
DTRG	1,0000	4,1429	2,0354	0,8600	2,0000	1,4100	1,6429	5,3214	2,3068
WSRG	1,3495	3,1324	1,7698	1,2000	3,1100	1,7600	1,2701	3,1441	1,7732

8.1

	A1			A2			Média		
LINR	0,8061	0,8339	0,9132	0,1600	0,0300	0,1800	0,4818	0,2985	0,5463
LSSR	1,2500	1,5625	1,2500	0,2500	0,0600	0,2500	0,7500	0,5625	0,7500
KNRG	1,1429	2,8889	1,6997	0,2400	0,1400	0,3800	0,6905	1,0754	1,0370
DTRG	0,5714	0,5714	0,7559	0,0000	0,0000	0,0000	0,2857	0,1429	0,3780
WSRG	1,5955	2,5792	1,6060	0,3300	0,1100	0,3300	0,9955	1,0014	1,0007

8.3

	A1			A2			Média		
LINR	1,1851	2,8053	1,6749	0,5100	0,4800	0,6900	0,6370	0,9776	0,9887
LSSR	1,1857	2,5471	1,5960	0,4800	0,4700	0,6900	0,6250	0,9056	0,9516
KNRG	1,3333	3,1746	1,7817	0,9000	1,1600	1,0800	0,7857	1,2976	1,1391
DTRG	1,1429	3,1429	1,7728	0,5700	0,5700	0,7600	0,9286	1,2500	1,1180
WSRG	1,1736	2,4839	1,5761	0,4400	0,4200	0,6500	0,5752	0,8388	0,9159

8.4

	A1			A2			Média		
LINR	1,4070	2,3603	1,5363	0,5000	0,5000	0,7100	0,9599	1,1394	1,0674
LSSR	1,8643	3,9654	1,9913	0,6300	0,6400	0,8000	1,2464	1,7985	1,3411
KNRG	1,2381	1,9683	1,4029	0,5200	0,4600	0,6800	0,7857	0,9008	0,9491
DTRG	1,4286	4,8571	2,2039	0,5700	0,8600	0,9300	1,0714	1,4643	1,2101
WSRG	1,7695	3,5869	1,8939	0,7300	0,7700	0,8800	1,2429	1,7818	1,3348

8.5

	A1			A2			Média		
LINR	1,5868	3,4500	1,8574	0,6300	1,0500	1,0200	1,0318	1,3200	1,1489
LSSR	1,5643	3,3311	1,8251	0,6800	1,1100	1,0500	1,0214	1,2511	1,1185

Continua na próxima página

University of North Texas

	Métricas								
Método	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
KNRG	1,3333	4,0000	2,0000	0,6200	0,9700	0,9800	0,8810	1,4087	1,1869
DTRG	2,0000	7,4286	2,7255	0,0000	0,0000	0,0000	1,6429	3,8214	1,9548
WSRG	1,5837	3,3846	1,8397	0,6600	1,0900	1,0400	1,0250	1,2573	1,1213

8.6

	A1			A2			Média		
LINR	0,9607	1,1437	1,0694	0,4800	0,4100	0,6400	0,7192	0,6822	0,8260
LSSR	1,2143	1,5529	1,2461	0,5700	0,5000	0,7100	0,8929	0,9196	0,9590
KNRG	0,5238	0,7778	0,8819	0,2900	0,3200	0,5600	0,4048	0,4643	0,6814
DTRG	0,7143	1,8571	1,3628	0,2900	0,5700	0,7600	0,7143	1,9286	1,3887
WSRG	1,2091	1,5179	1,2320	0,5400	0,4800	0,6900	0,8895	0,9019	0,9497

8.7

	A1			A2			Média		
LINR	0,6639	0,6307	0,7942	0,3800	0,2300	0,4800	0,4503	0,3396	0,5828
LSSR	0,8571	0,8571	0,9258	0,3900	0,2100	0,4500	0,5893	0,4085	0,6391
KNRG	0,9524	1,2698	1,1269	0,4800	0,3200	0,5600	0,7143	0,6508	0,8067
DTRG	0,8571	1,4286	1,1952	0,4300	0,4300	0,6500	0,5000	0,5357	0,7319
WSRG	0,9061	1,1020	1,0498	0,3000	0,2700	0,5200	0,5861	0,5359	0,7320

9.1

	A1			A2			Média		
LINR	0,2602	0,0736	0,2713	0,0300	0,0000	0,0500	0,1567	0,0266	0,1631
LSSR	0,2500	0,0625	0,2500	0,0500	0,0000	0,0500	0,1500	0,0225	0,1500
KNRG	0,0476	0,0159	0,1260	0,0500	0,0200	0,1300	0,0476	0,0159	0,1260
DTRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,2143	0,3214	0,5669
WSRG	0,3192	0,1024	0,3201	0,0400	0,0000	0,0400	0,1766	0,0313	0,1769

9.2

	A1			A2			Média		
LINR	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
LSSR	0,1500	0,0225	0,1500	0,1000	0,0100	0,1000	0,1250	0,0156	0,1250
KNRG	1,0000	1,0000	1,0000	0,6700	0,4400	0,6700	0,8333	0,6944	0,8333
DTRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
WSRG	0,3995	0,1596	0,3995	0,2600	0,0700	0,2600	0,3389	0,1148	0,3389

9.3

	A1			A2			Média		
LINR	1,0469	1,4897	1,2205	0,3400	0,2500	0,5000	0,6280	0,6398	0,7999
LSSR	1,2429	2,0100	1,4177	0,3500	0,2200	0,4700	0,7321	0,7835	0,8851
KNRG	0,6190	0,6508	0,8067	0,4300	0,2400	0,4900	0,4762	0,3333	0,5774

Continua na próxima página

University of North Texas									
					Métricas				
Método	MAE	MSE	RMSE	MAE	MSE	RMSE	MAE	MSE	RMSE
DTRG	1,8571	5,8571	2,4202	0,4300	0,4300	0,6500	1,2143	2,3929	1,5469
WSRG	1,3984	2,3493	1,5327	0,3200	0,2400	0,4900	0,8118	0,9299	0,9643

9.4

	A1			A2			Média		
LINR	1,1510	1,5710	1,2534	0,5000	0,2800	0,5300	0,7113	0,6720	0,8198
LSSR	1,0714	1,3929	1,1802	0,5400	0,3800	0,6200	0,7500	0,7082	0,8416
KNRG	0,5714	0,7302	0,8545	0,4300	0,3700	0,6000	0,5000	0,3611	0,6009
DTRG	0,5714	0,8571	0,9258	0,2900	0,2900	0,5300	0,6429	1,0357	1,0177
WSRG	1,0551	1,3512	1,1624	0,5400	0,3800	0,6100	0,7466	0,7092	0,8422

9.5

	A1			A2			Média		
LINR	0,9757	2,8749	1,6956	0,3400	0,4900	0,7000	0,6421	1,4386	1,1994
LSSR	1,0929	3,0739	1,7533	0,3600	0,5200	0,7200	0,7107	1,5199	1,2328
KNRG	0,9048	2,9048	1,7043	0,2900	0,5700	0,7600	0,5952	1,4881	1,2199
DTRG	0,7143	2,4286	1,5584	0,2900	0,5700	0,7600	0,5714	0,9286	0,9636
WSRG	1,0554	3,0868	1,7569	0,3200	0,5400	0,7300	0,6893	1,5336	1,2384

9.6

	A1			A2			Média		
LINR	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
LSSR	0,5000	0,2500	0,5000	0,2000	0,0400	0,2000	0,3500	0,1225	0,3500
KNRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
DTRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
WSRG	0,2706	0,0732	0,2706	0,1700	0,0300	0,1700	0,2212	0,0489	0,2212

9.7

	A1			A2			Média		
LINR	0,1572	0,0432	0,2079	0,1200	0,0300	0,1600	0,1297	0,0295	0,1716
LSSR	0,2000	0,0400	0,2000	0,1500	0,0200	0,1500	0,1750	0,0306	0,1750
KNRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
DTRG	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
WSRG	0,0365	0,0013	0,0366	0,0500	0,0000	0,0500	0,0428	0,0019	0,0433

Tabela 17 – Resultados individuais para as atividades da base de dados da UNT.

Última página