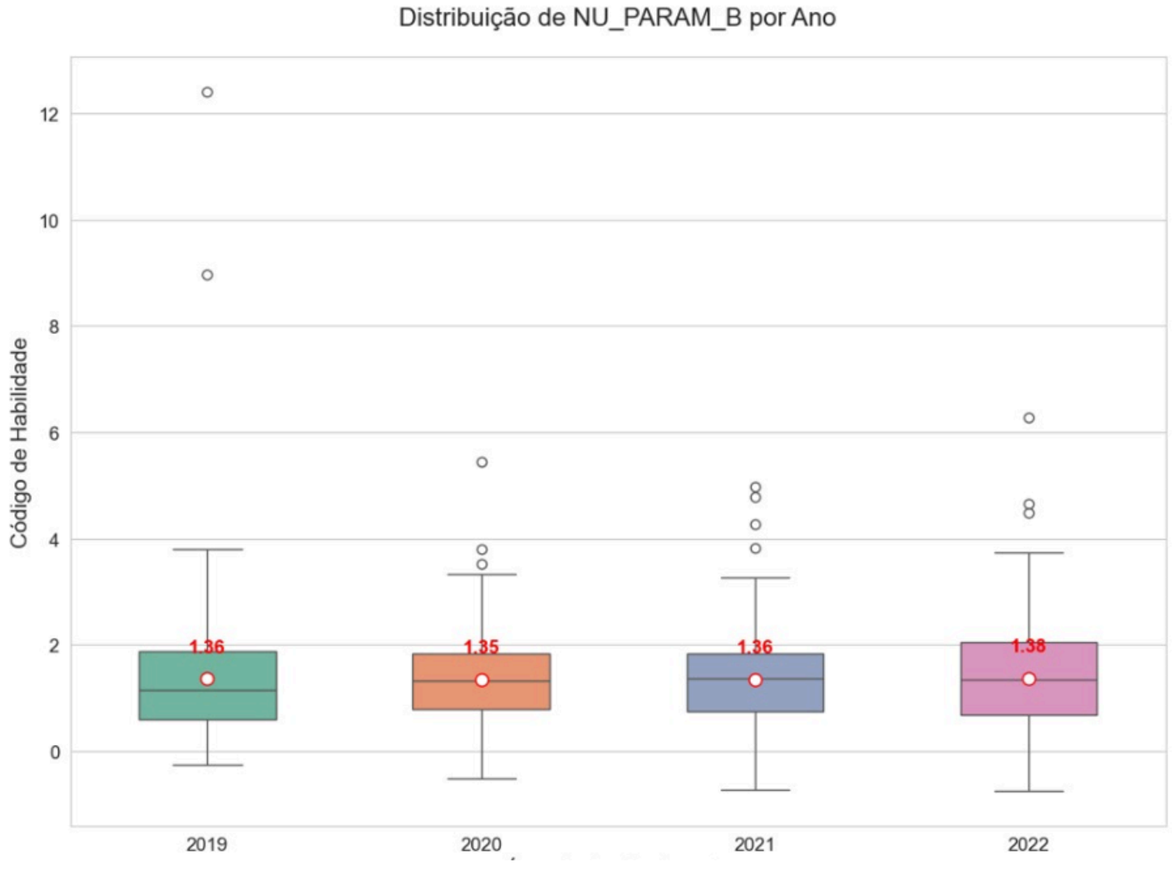
PREDIÇÃO DE DIFICULDADE DE QUESTÕES DO ENEM

INTRODUÇÃO

O ENEM É UMA DAS PRINCIPAIS AVALIAÇÕES EDUCACIONAIS DO BRASIL. CADA QUESTÃO DA PROVA POSSUI UM PARÂMETRO DE DIFICULDADE (B), ESTIMADO APÓS A APLICAÇÃO COM BASE NA TEORIA DA RESPOSTA AO ITEM (TRI).NESTE TRABALHO, BUSCAMOS PREVER ESSE PARÂMETRO DE DIFICULDADE ANTES DA APLICAÇÃO DA PROVA, UTILIZANDO FEATURES EXTRAÍDAS AUTOMATICAMENTE DO TEXTO DAS QUESTÕES, COMBINADAS COM OUTRAS VARIÁVEIS DISPONÍVEIS NO BANCO DE DADOS DO EXAME. ESSA PREDIÇÃO PODE APOIAR A CONSTRUÇÃO DE PROVAS MAIS EQUILIBRADAS E EFICIENTES.

DADOS

FORAM UTILIZADOS OS ENEMS DE 2019, 2020 E 2022 PARA TREINO E O ENEM 2023 COMO TESTE. OBSERVOU-SE QUE TODAS AS PROVAS APRESENTARAM MENOS DE 180 QUESTÕES VÁLIDAS, INDICANDO O ANULAMENTO DE ITENS EM TODOS OS ANOS ANALISADOS (ENTRE 173 E 178 QUESTÕES).



METODOLOGIA

- **COLETA DOS DADOS: AS PROVAS DO ENEM FORAM CONVERTIDAS DE PDF PARA JSON COM O ENEM EXTRACTOR, PRESERVANDO O ENUNCIADO, O PARÂMETRO DE DIFICULDADE (B) E A PRESENÇA DE IMAGEM.**
- **PRÉ-PROCESSAMENTO: SELECIONAMOS APENAS AS QUESTÕES COM TEXTO COMPLETO E PARÂMETRO DE DIFICULDADE DISPONÍVEL; UNIMOS OS JSONS AOS MICRODADOS DO INEP E REMOVEMOS ITENS COM DADOS FALTANTES.**
- **EXTRAÇÃO DE FEATURES: UTILIZAMOS A BIBLIOTECA AIBOXLAB/NLP PARA EXTRAIR CARACTERÍSTICAS LINGUÍSTICAS DOS ENUNCIADOS.**

RESULTADOS

- FORAM TESTADOS SEIS MODELOS DE REGRESSÃO: SVR, RIDGE, LASSO, DECISION TREE, RANDOM FOREST E LINEAR REGRESSION.
- O MELHOR DESEMPENHO FOI O DO SVR, COM R^2 DE 0.0566 NO TESTE E 0.0787 NA PROVA DE 2023, INDICANDO BAIXA CAPACIDADE PREDITIVA.
- A COESÃO REFERENCIAL FOI A FEATURE COM MAIOR CORRELAÇÃO POSITIVA (≈0.33), ENQUANTO A DIVERSIDADE LEXICAL TEVE A MAIOR NEGATIVA (≈−0.26).
- UTILIZAR APENAS AS 10 FEATURES MAIS CORRELACIONADAS NÃO MELHOROU OS RESULTADOS.
- OS RESULTADOS SUGEREM QUE MÉTRICAS LINGUÍSTICAS SUPERFICIAIS NÃO SÃO SUFICIENTES PARA ESTIMAR A DIFICULDADE DAS QUESTÕES.