

**UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE**

**CENTRO DE CIÊNCIAS EXATAS E DA TERRA**

**CURSO DE ESTATÍSTICA.**

**ANA LUZIELMA DIAS CAMPOS**

**JAYLHANE VELOSO NUNES**

**RAIANNY DA SILVA SOARES**

**ANÁLISE DE AVALIAÇÕES DE LIVROS USANDO A TÉCNICA XGBOOST**

**NATAL/RN**

**DEZEMBRO/2021SUMÁRIO**

| **1**  **1.1** | **INTRODUÇÃO**........................................................................................  **Objetivo geral**…………………….........……………………………………… | 3  3 |
| --- | --- | --- |
| **2** | **DESCRIÇÃO**........................................................................................... | 4 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**1 INTRODUÇÃO**

Quando se está procurando uma nova leitura, uma das coisas que pode ser observada é a avaliação do livro. De acordo com ela, pode-se ter uma ideia inicial se o livro é bom, já que muitas pessoas deram uma nota alta de avaliação.

Pensando sobre isso, levantamos o questionamento: seria possível predizer se um livro é bom sem ter acesso a nota da avaliação? Dessa forma pensamos em realizar um modelo para classificar a avaliação de um livro.

Para esta tarefa utilizaremos o seguinte conjunto de dados: [Goodreads-books | Kaggle](https://www.kaggle.com/jealousleopard/goodreadsbooks) e como inspiração para construção do modelo utilizaremos o seguinte guia: [**Tune xgboost models with early stopping to predict shelter animal status | Julia Silge**](https://juliasilge.com/blog/shelter-animals/)

Assim, uma das possibilidades que foram pensadas é fazer uma categorização das avaliações dos livros em “Ruim”, “Bom” e ''Ótimo”, considerando respectivamente os intervalos de nota abaixo de 2.5, entre 2.5 a 3.5 e acima de 3.5, e a partir daí predizer a avaliação dos livros utilizando o XGboost. O critério de intervalo para as categorias das notas foi definido subjetivamente ao acaso entre as participantes do grupo.

Além disso, como o objetivo é classificar os livros sem olhar as avaliações, as notas não farão parte do modelo, elas serão utilizadas apenas para criar as categorias e estamos supondo que de alguma forma as variáveis como número de páginas, idade do livro, editora, quantidade de notas de avaliações e quantidade de avaliações escritas estão relacionadas com a avaliação do livro.

Ficamos curiosas com outros pontos que podemos incluir no modelo, tais como: será que o mês em que o livro foi lançado influencia? Ou o gênero do autor?

Dado que temos a variável com o nome dos autores, podemos predizer o gênero por meio de pacotes como [Gender](https://cran.r-project.org/web/packages/gender/gender.pdf) e [genderdata](https://github.com/lmullen/genderdata). Um empecilho para usar esses pacotes é o fato de que alguns autores constam no conjunto de dados com nomes abreviados, por isso essa ideia ainda está em análise.

Dessa forma, podemos definir os objetivos do trabalho como:

**1.1 Objetivo geral**

Criar um modelo de classificação utilizando o XGBoost.

**2 DESCRIÇÃO**

O banco de dados tem informações de 12 variáveis, que serão mostradas a seguir:

| Variável | Descrição |
| --- | --- |
| bookID | Um número específico de identificação para cada livro |
| title | O título do livro |
| authors | Nome(s) do/os Autor(a)/Autores(as) |
| average\_rating | Avaliação média do livro |
| isbn | Outro tipo de identificação exclusiva para o livro (International Standard Book Number.) |
| isbn13 | ISBN de 13 dígitos para identificação do livro |
| language\_code | Idioma principal(de origem?) do livro |
| num\_pages | Número de páginas |
| ratings\_count | Número total de avaliações que o livro teve |
| text\_reviews\_count | Número total de resenhas feitas no livro |
| publication\_date | Data da primeira publicação |
| publisher | Nome do/a editor/a |