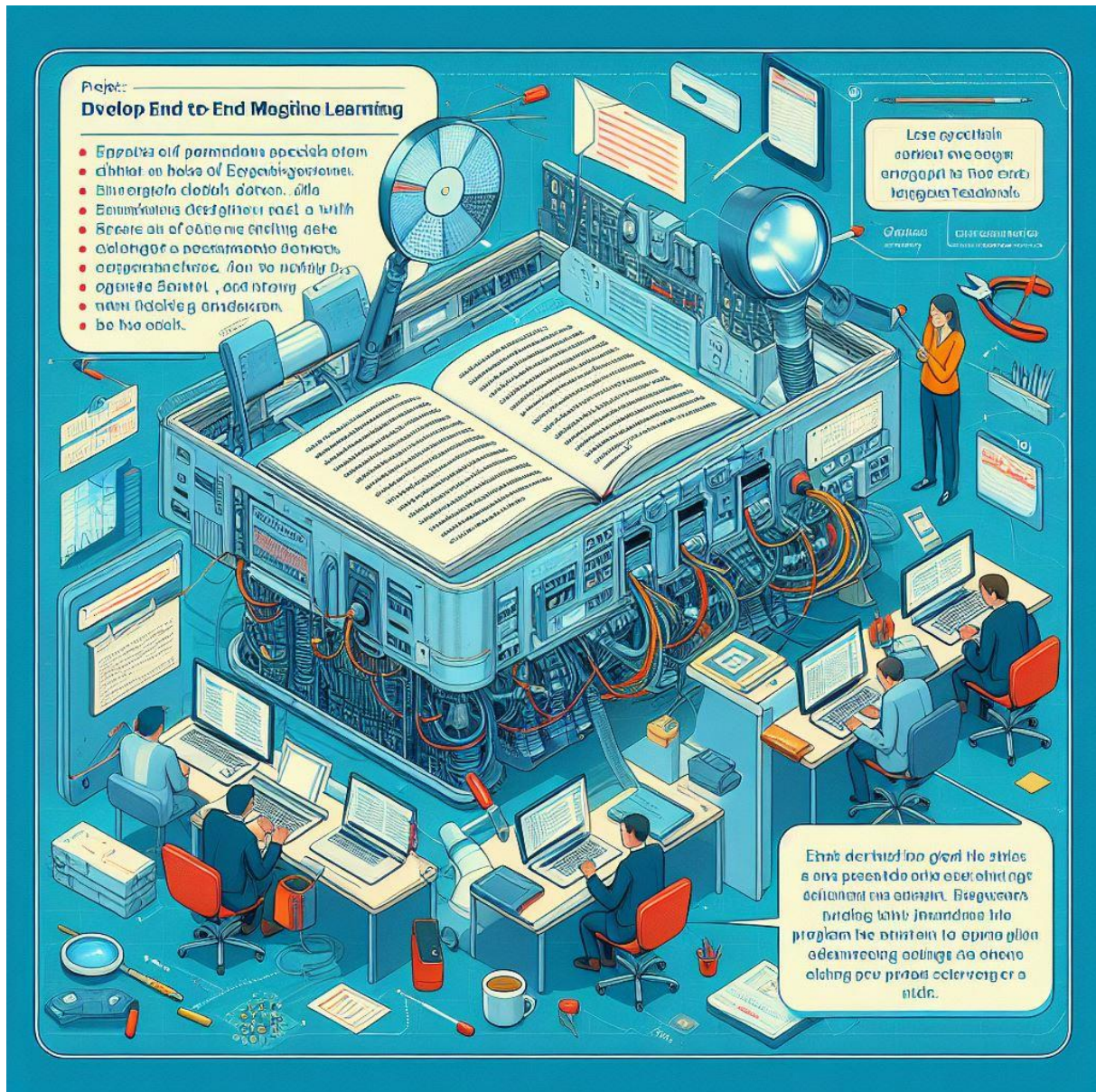


## Groupe 6 : Martial Kouassi, Naïma Taïleb, Cedric Djivo, Michael Rovia

Github project : <https://github.com/cedricdji/easy-sailing-language-training-machine-learning-project-dsti>

### Project: Develop an end-to-end Machine Learning Pipeline

Instructor: Assan Sanogo



## 1. Introduction/Business Understanding

The realm of natural language processing (NLP) offers a vast landscape of opportunities for machine learning applications. In this project, our objective revolves around developing a robust end-to-end pipeline for

predicting the quality of English essays. The dataset comprises over 7000 essays graded by English specialists, simulating a real-world scenario where assessing the proficiency of English language learners is essential.

### Project Overview

Our project centers on constructing a comprehensive machine learning pipeline capable of processing essays and providing a grade indicative of the writer's English proficiency. Leveraging the interdisciplinary nature of our team, consisting of Data Analysts, Data Scientists, and Data Engineers, we aim to harness diverse skill sets for optimal results.

### Project Objectives














The primary goal of this project is to train a model using the provided dataset to predict the rating of an essay. We intend to encapsulate the entire process within a Jupyter Notebook, incorporating stages such as exploratory data analysis, feature engineering, model training, evaluation, and potentially deployment.

### Problem Reframing

Upon careful consideration of the dataset, if challenges arise, we are open to reframing the problem as a classification task, categorizing essays into levels of proficiency, such as poor, average, or great.

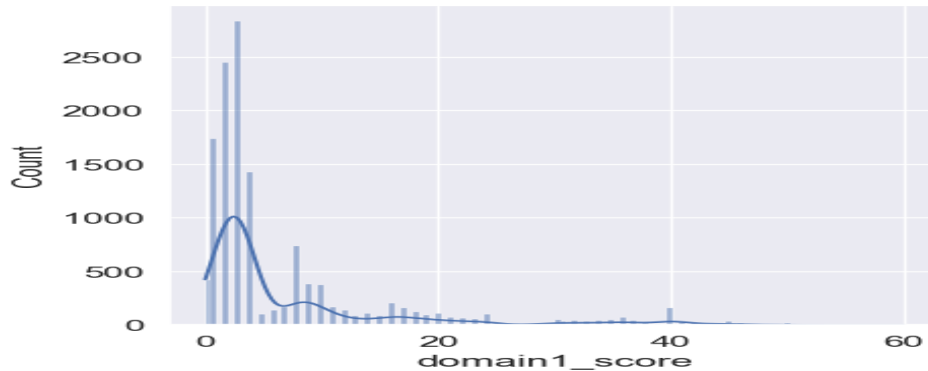
## 2. Data description

The dataset consists of essays graded by English specialists, providing a rich resource for training and evaluation.

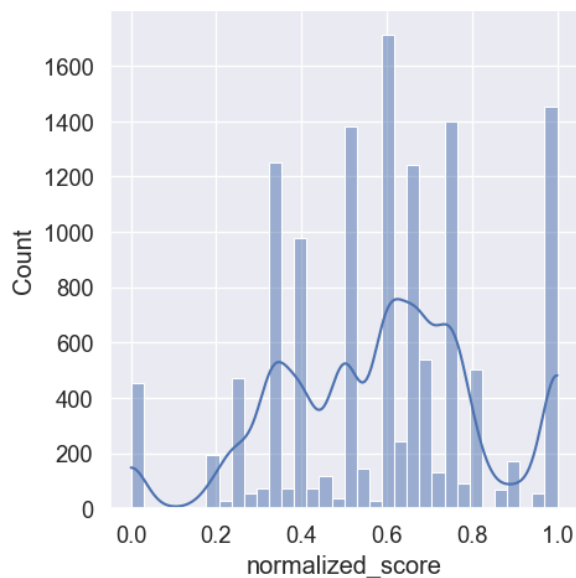
 Essay_Set_Descriptions	Dossier compressé	213 Ko	
 test_set.tsv	Fichier TSV	1 640 Ko	
 Training_Materials	Dossier compressé	56 292 Ko	
 training_set_rel3	Feuille de calcul Microsoft Exc...	5 623 Ko	
 training_set_rel3	Feuille de calcul Microsoft Excel	6 353 Ko	
 training_set_rel3.tsv	Fichier TSV	4 981 Ko	
 valid_sample_submission_1_column	Fichier CSV Microsoft Excel	3 Ko	
 valid_sample_submission_1_column_no_header	Fichier CSV Microsoft Excel	3 Ko	
 valid_sample_submission_2_column	Fichier CSV Microsoft Excel	13 Ko	
 valid_sample_submission_5_column	Fichier CSV Microsoft Excel	23 Ko	
 valid_set	Feuille de calcul Microsoft Exc...	1 834 Ko	
 valid_set	Feuille de calcul Microsoft Excel	2 062 Ko	
 valid_set.tsv	Fichier TSV	1 640 Ko	

### 3. Data Preprocessing

Following the importation of datasets, we meticulously assessed them for missing values, adhering to a stringent threshold that dictated the dropping of columns with missing values exceeding 80%. Furthermore, our in-depth analysis uncovered biases within the dataset, notably a concentration of grades between 0 and 10.

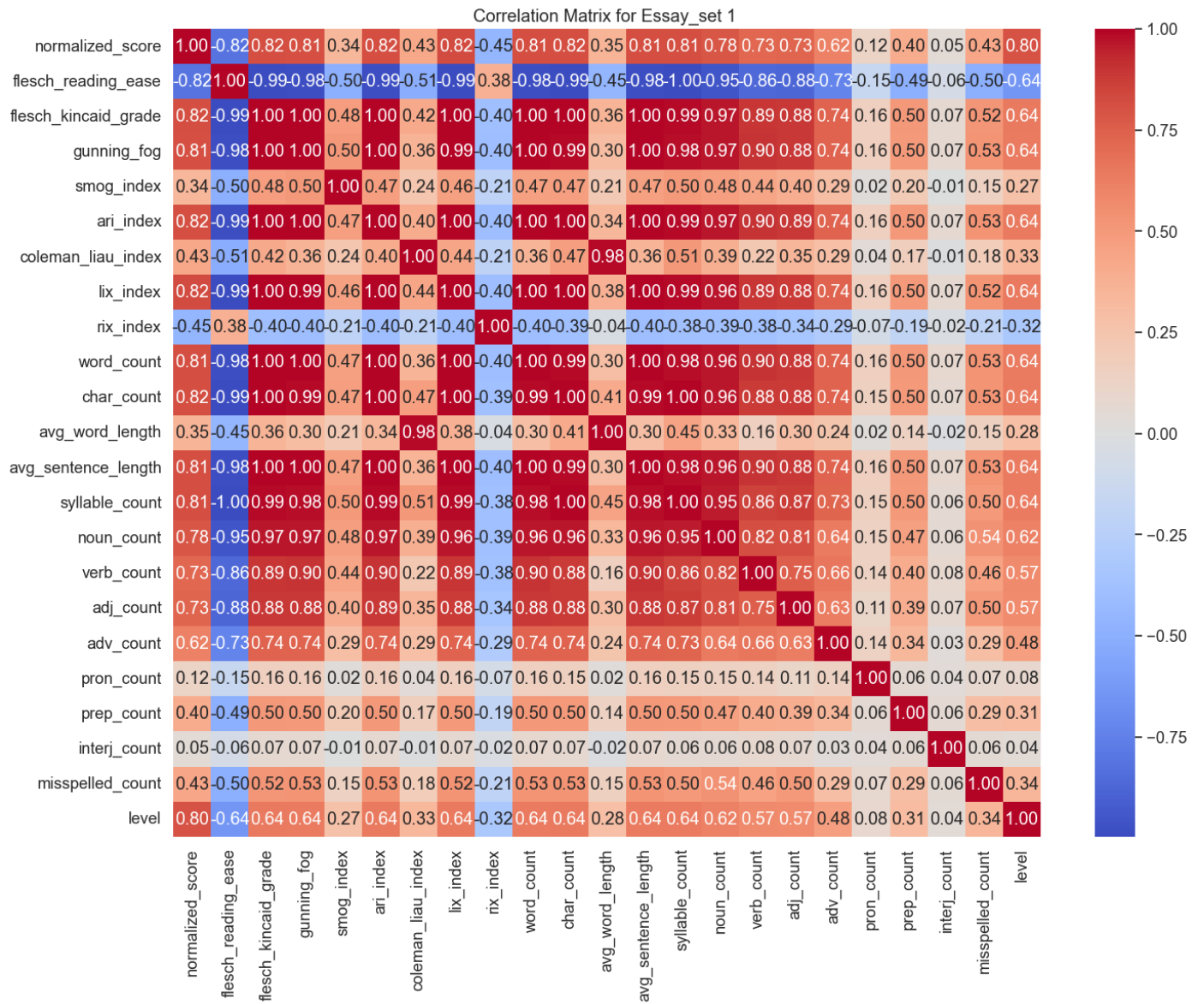


To rectify this imbalance, we judiciously applied a normalization function, thereby achieving a more equitable grade distribution across the spectrum.



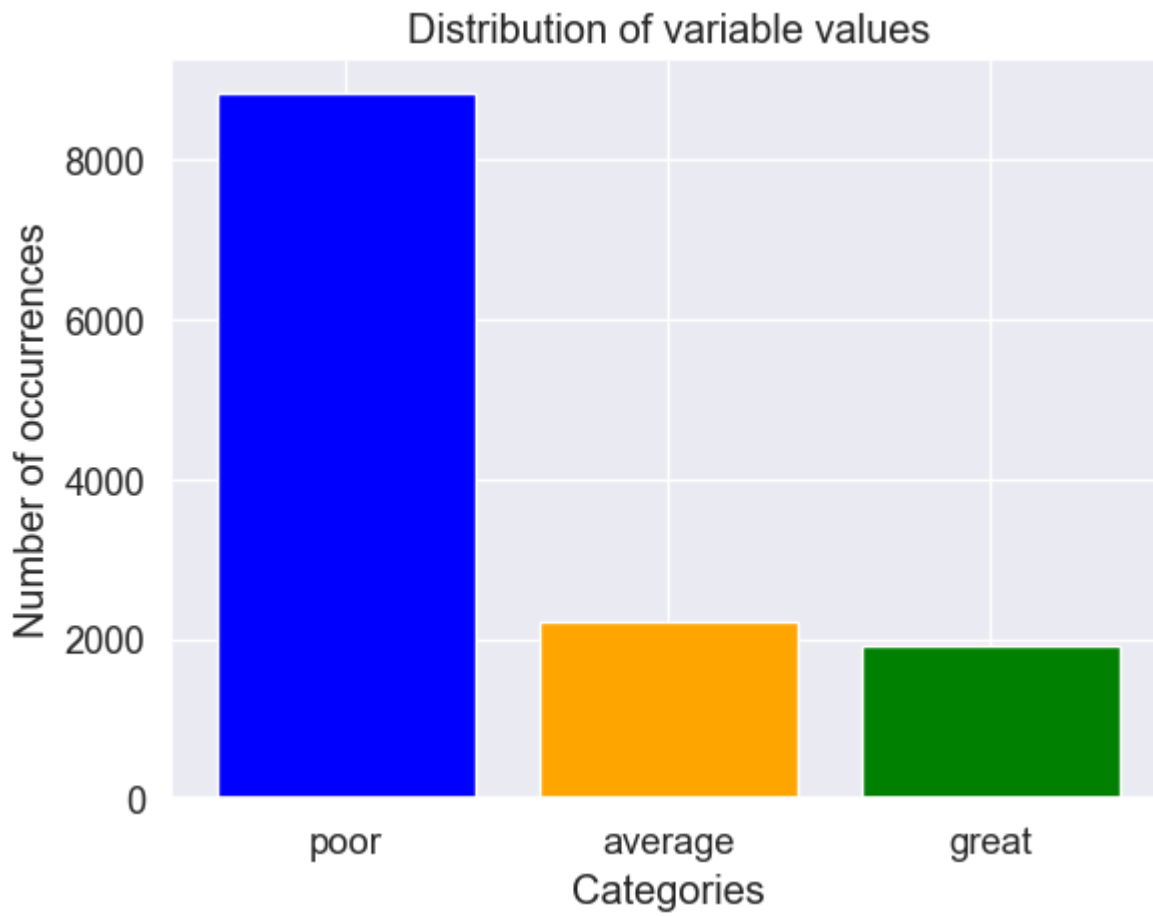
Additionally, we delved into the intricacies of text cleaning and feature engineering, where we meticulously crafted features to encapsulate text complexity and quality. This phase was augmented by the creation of correlation matrices for each essay type, serving as invaluable guides to inform subsequent modeling decisions.



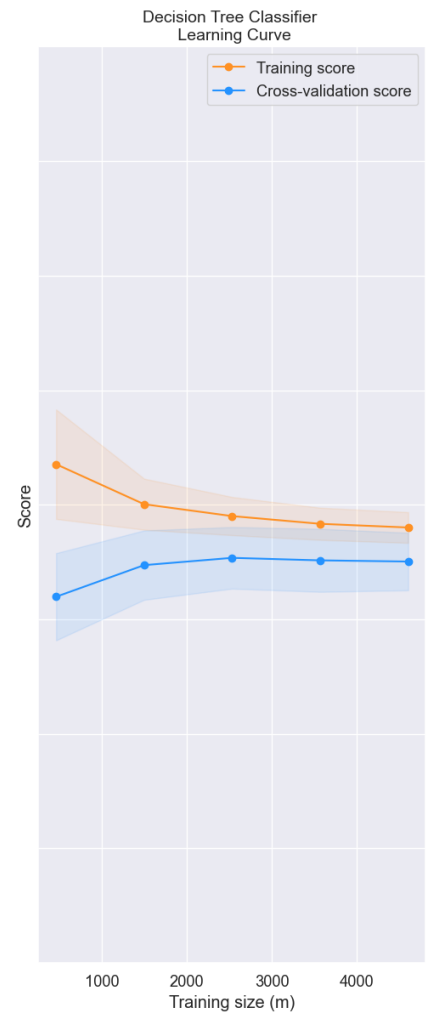
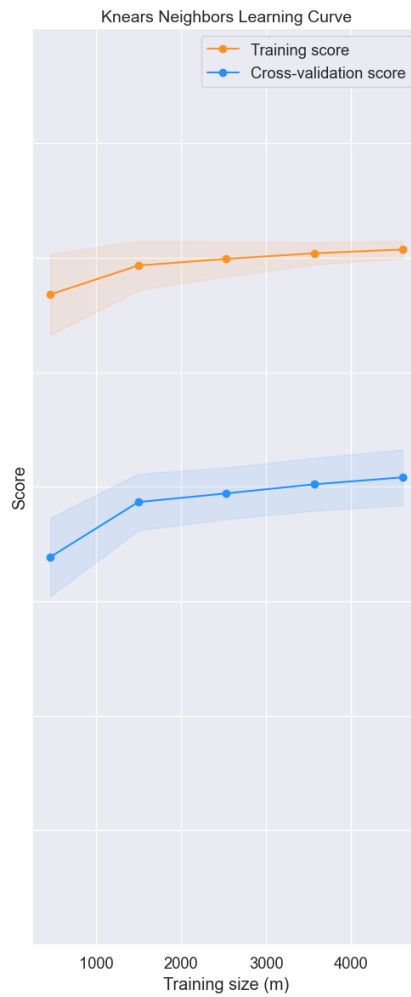
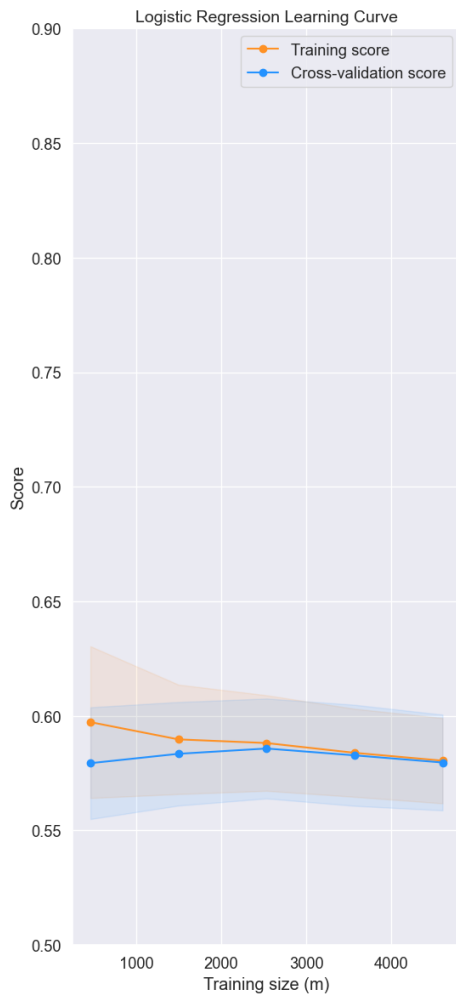


## 4. Modeling and Evaluation

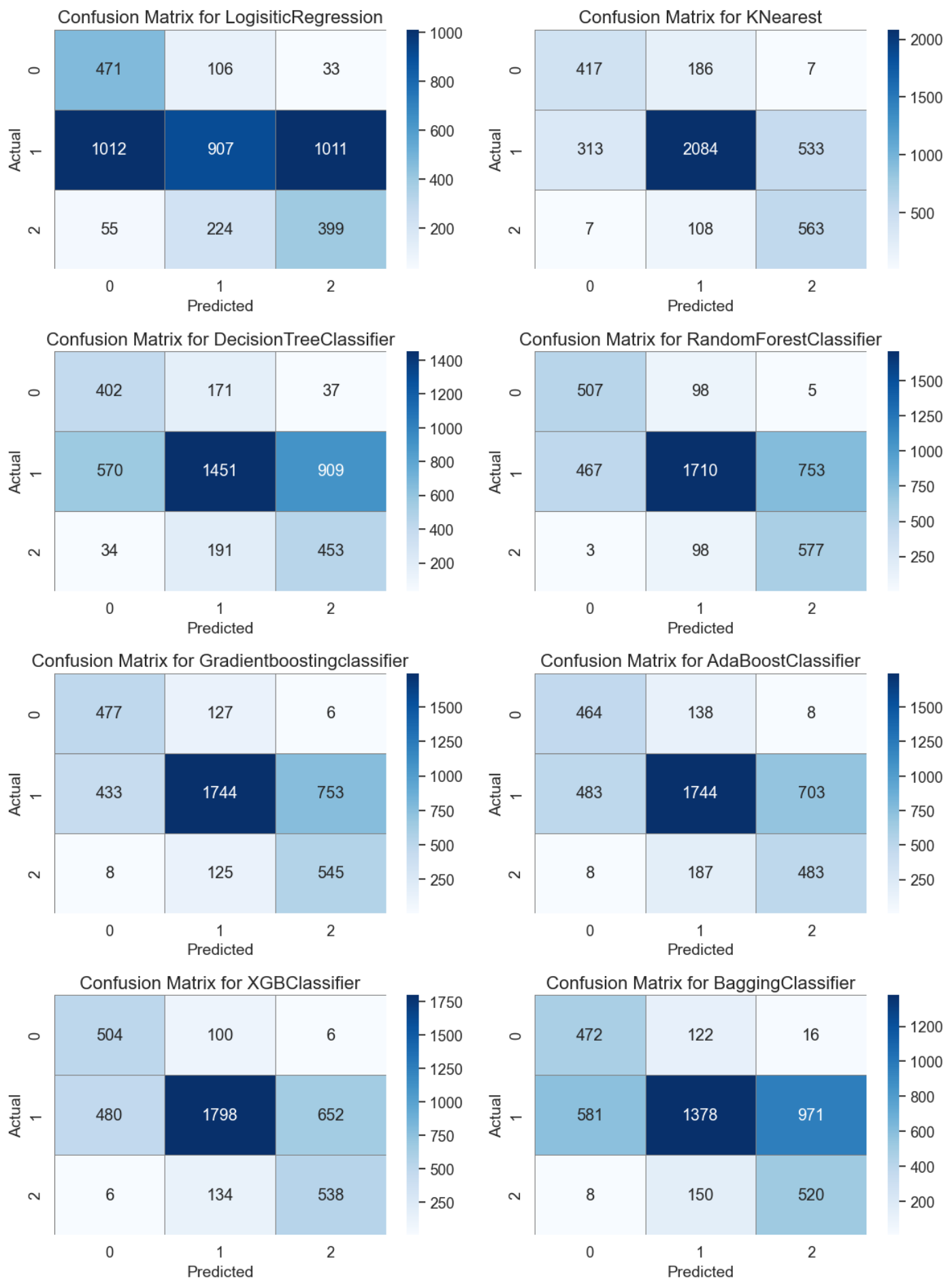
Our modeling endeavors were marked by meticulousness and rigor, as we sought to construct a validation set that faithfully mirrored the transformations applied to the training set. This meticulous approach paved the way for fair and equitable model evaluation, a crucial aspect in the pursuit of model excellence. Moreover, we undertook the arduous task of assigning English proficiency levels (poor, average, great) to each essay, a critical step preceding the training of various classifiers.



Through rigorous cross-validation techniques, we endeavored to optimize model performance, culminating in a meticulous evaluation of classifier efficacy.



# confusion matrix



Our meticulous analysis identified the RandomForestClassifier as the most potent performer, boasting an impressive precision score of 66%.

## 5. Conclusion

In conclusion, our journey through this machine learning project has been characterized by meticulousness, diligence, and a relentless pursuit of excellence. While our model has demonstrated commendable performance, achieving a precision score of 66%, we recognize that the pursuit of perfection is an ongoing endeavor. To this end, we envision future enhancements that span the breadth of feature engineering, classifier refinement, and model interpretability. By embracing advanced techniques such as Count n-grams and leveraging sophisticated text vectorizers, we anticipate unlocking new realms of performance excellence. Furthermore, we envisage delving into the intricacies of neural network architectures and employing feature selection methodologies to augment model interpretability and mitigate overfitting. In essence, this project serves as a testament to our commitment to pushing the boundaries of machine learning excellence and charting new frontiers in English proficiency classification.