Evaluating IELTS writing: A Machine Learning application

DANG MANH CUONG 20214949
TRAN MINH TUAN 20214978
TRINH TRUONG GIANG 20214958

September 23, 2023

Contents

Abstract

With the growing social development and international communication, more and more Vietnamese take the IELTS test as the basis of study and work. As a result, IELTS test preparation is a booming business in recent years, given the test's difficulty and its high-stake nature. One component that is often received the most attention from prospective candidates is Writing. Candidates often find preparing for this component challenging because of the inherent difficulty assessing one's writing skills on their own and the fact that the assessment rubrics made public by official IELTS partners are extremely ambiguous for self-studying. In this project, we attempt to create a system that can assess candidates' Writing answers without human raters' presence, taking advantage of various Machine Learning algorithms in accordance with established measurement and evaluation indicators. While the performance results are less than stellar, the project can be further developed into a platform to provide candidates who self-study for the test with necessary support in order to prepare for the Writing component. In addition, this system can also substitute human markers or support human markers with marking Writing responses, ensuring complete impartiality and objectivity in the assessment of this high-stake test.

Acknowledgement

We would like to express our gratitude to Assoc. Prof. Than Quang Khoat for giving us an opportunity to work together on a practical topic in the Machine Learning course this semester. The knowledge we gained from the course has helped us in navigating through the problem and delivering the right approach to the problem.

Chapter 1

Introduction

The International English Language Testing System (IELTS) is widely considered as the most popular English language test for migration and higher education[1]; particularly, in Vietnam, the demand of the test has skyrocketed in recent years[2]. In addition to the aforementioned purposes, IELTS test results in Vietnam are also considered a reliable proof of English language competence and a requirement to gain admission into many Vietnamese universitites[2].

In order to receive an IELTS test result, a candidate must complete 4 components: Listening, Reading, Writing and Speaking[3]. While it's possible for any potential candidate to objectively assess the Listening and Reading skills on their own (IELTS practice resources from official sources provide exact answers for listening and reading sections [4]). Writing, alongside Speaking, are the components for which band score is awarded subjectively by human examiners on a scale from 1 to 9 - which means personal/professional opinion of examiners may affect the scoring of the component. As a result, it is difficult for a candidate to assess their Writing skill on their own. IELTS partners do provide a so-called "public version" of the Writing assessment rubrics[5][6] - which may supposedly support candidates with their preparation; however, the wording in those rubrics are ambigious, which further compounds the challenges that candidates are already facing.

The only remaining option for candidates is sample answers of various band scores[4] - "implementations" of the assessment rubrics - which are available in various preparation materials available freely or by paying.

This is where we see the potential of applying machine learning into this problem as a multiclass classfication problem. By utilizing machine learning, we can derive the inner workings of IELTS writing assessment scheme based on such scripts without access to proper marking rubrics.

As information on IELTS marking rubrics is sparse, this project's approach will be somewhat unusual. Instead of attempting to recreate features similar to IELTS marking criteria, we explores various ways in which candidate texts can be analyzed based on standard features that are commonly used in text classification to see if a band score can be reliably produced from such features. Chapter 2 presents the methodology by which the study is conducted. Chapter 3 describes the evaluation results, discussing them and identifies directions for future endeavors.

Why writing?

Writing is an important language skill and one of the main indicators of one's ability to use a language. It is not only an important way to communicate ideas, emotional expressions and cultural exchanges, but also reflects the language ability, communication ability and thinking level in that language. Therefore, in the context of English, English writing ability is an important index of English professional ability and an important aspect of one's English professional training. That's why it is important to focus on improving one's writing. Nevertheless, writing is also quite difficult to master and unappealing to many people.

Chapter 2

Methodology

2.1 Dataset description

Despite the popularity of IELTS, finding a dataset of IELTS writing scripts is challenging. While answers for Writing questions are readily available on the Internet, there is no publicly available central database of writing responses for IELTS. Therefore, we have spent a significant amount of time (across 2 months: May 2023 and June 2023) gathering data from a multitude of sources It is worth noting that sources from Cambridge English (the sole producer of IELTS tests) are particularly emphasized, as it is the most authoritative[7] source of Writing scripts publicly available (freely or by paying). In addition, there are many reputable publishers and IELTS ex-examiners (e.g. Collins or ielts-simon website) also publish preparation materials for IELTS - which are also used as a source for the dataset. Because some private sources were involved in the collection of our dataset, unfortunately, we cannot disclose the full list of sources that we used. This is to protect the sources - any disclosure in a document in the academic context might seriously hamper their reputation.

Regarding the scores given in the dataset, they are given by human raters on scale from 1 to 9. While efforts have been made to construct a dataset that covers the broadest range of scoring that is possible, responses corresponding to the scores at both extreme ends (below 4 and over 8) are a rare occurrence. This is supported by official IELTS test statistics[8]. However, we understand that this will mean the model cannot make a prediction in these edge cases.

2.2 Data preprocessing

Assessing a particular piece of writing involves analyzing the nuances of writing; therefore, extensive data preprocessing in this case is unnecessary. For this dataset, in particular, preprocessing is limited to converting the texts to lowercase. After that, each corpus is transformed (VSM) using the CountVectorizer function in the SKlearn package to extract the features. CountVectorizer is often used to convert a collection of text documents to a matrix of token counts.

```
vectorizer = CountVectorizer()

X_bow = pd.DataFrame(X_bow.toarray(),
```

```
columns = vectorizer.get_feature_names_out())
```

```
(1000, 9732)
```

As shown above, we ended up with 9732 features after converting the text into vectors.

2.3 Feature selection

While the previous steps have produced a significant number of features to be utilized during the training phase, it is clear that we need to explore other features that may be useful to improve the performance of the model. For starters, a feature is an individual measurable property or characteristic of a phenomenon.[9] Choosing informative, discriminating and independent features is a crucial element of effective algorithms in pattern recognition, classification and regression.

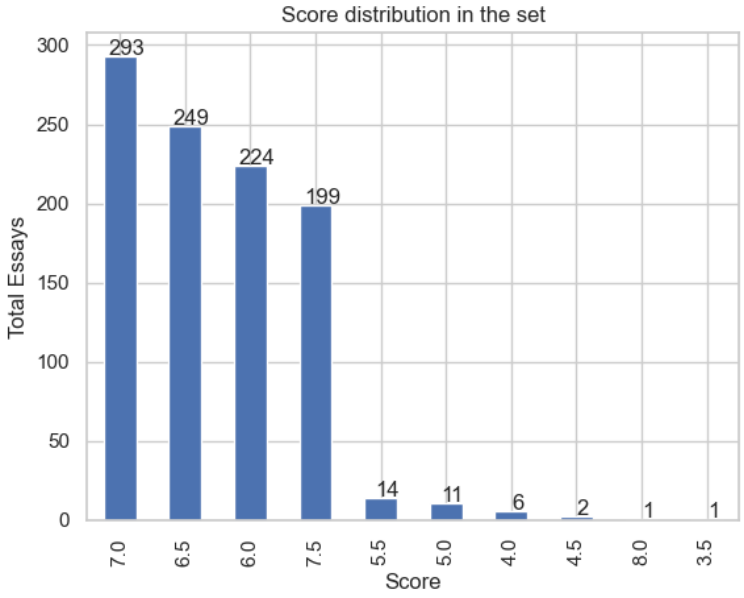
To avoid making things complicated, we have decided that a small set of features will be used. They cover basic statistics of an essay text: word length, character length, average word length, sentiment polarity and subjectivity. Sentiment-based features come from the TextBlob package.

2.4 Exploratory data analysis

Data information

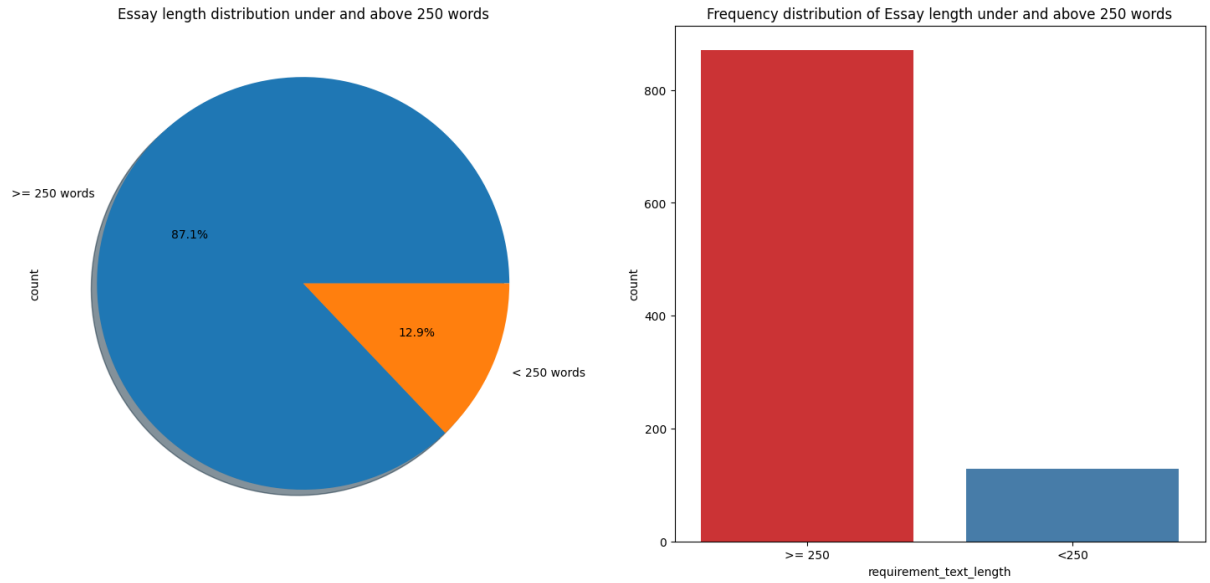
```
RangeIndex: 1000 entries, 0 to 999
Data columns (total 8 columns)
...
```

This dataset is small (1000 observations). There is no missing value. The chart below (created with Matplotlib) show how the scores are distributed in the dataset.



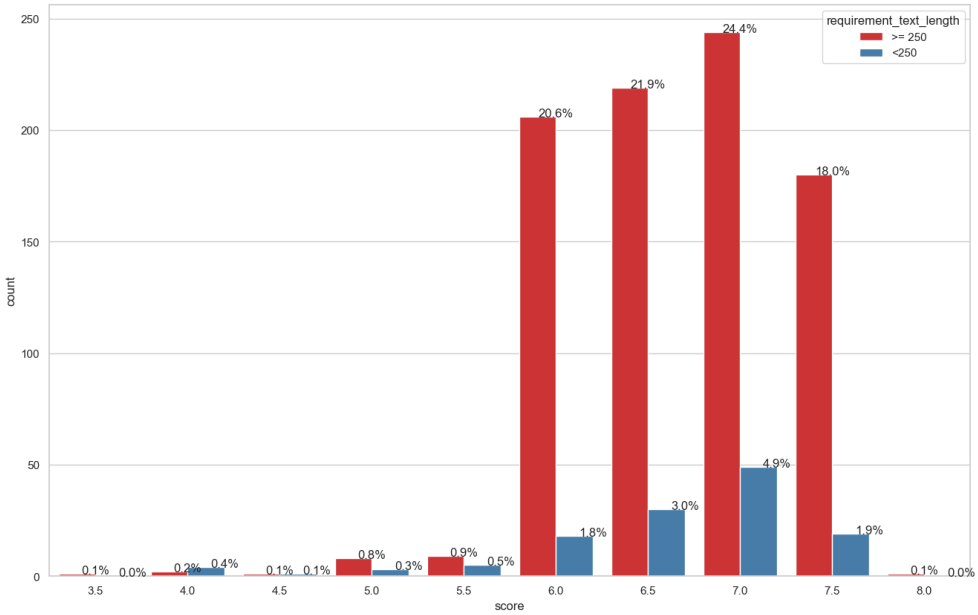
It is clear that the dataset is highly imbalanced, with the dataset covering the range from 4.0 to 8.0, with particular emphasis on band 6 to 7. This imbalance is expected given that the data is randomly collected from multiple sources. This will cause some unwanted behaviors during the training phase, as we will see in the coming sections.

Another thing we want to look at is how many essays in our dataset meet the minimum word length for the IELTS Writing task in question. The minimum word length for an Academic Writing Task 2 response is 250 words[3]. In actual test, scripts under the aforementioned word length will be penalized[3]); therefore, it is interesting to see if most candidates adhere to the word length limit imposed.



The graph above confirms our assumption: The majority of essays do comply with the word length requirement. This is not surprising, given the word length requirement is one of the easier requirement to satisfy in IELTS Writing.

Now we will take a look at the score distribution again, this time from the word length point of view. It is worth mentioning again that the score that we collected are from human raters, who have taken account of underlength in the final scoring. Therefore, we can see if the texts of acceptable score (¿= 6) are also with the acceptable word length or not. The answer is yes, according to this figure:

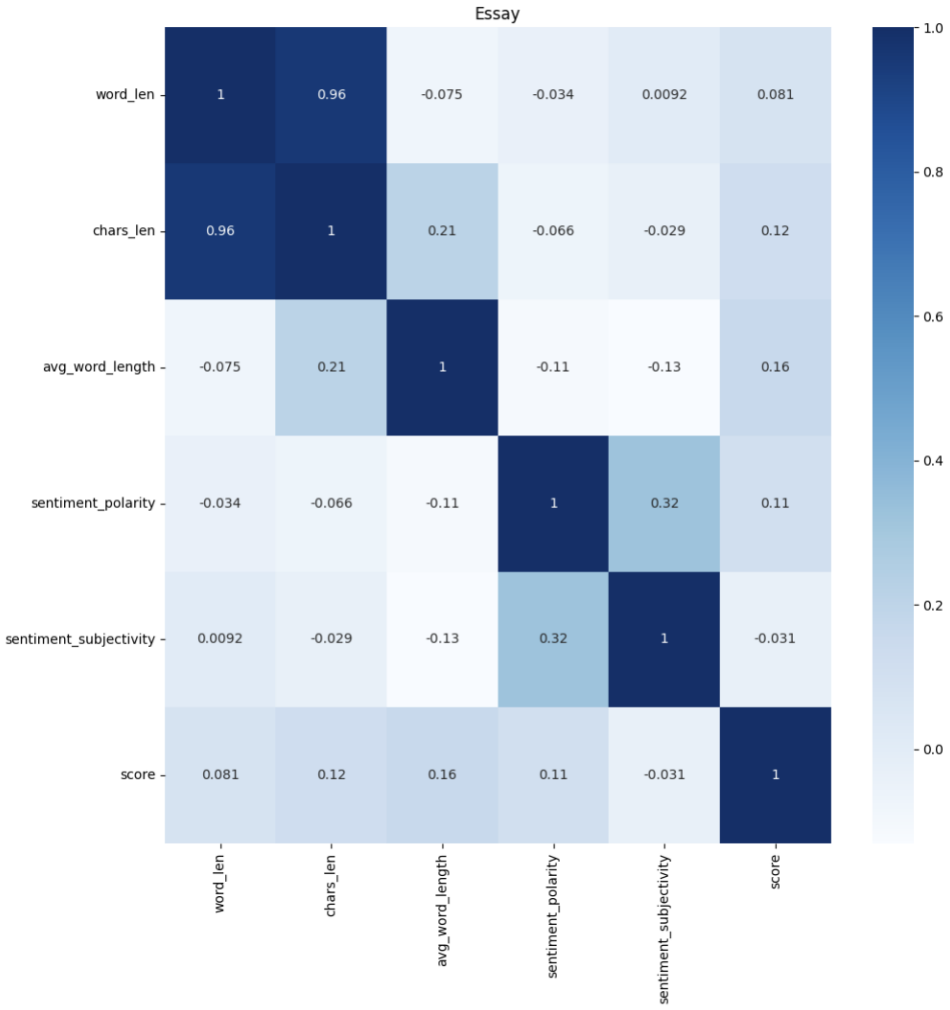


As the data for scores outside 6-7.5 are negligible, we will only consider the scores in the range mentioned. Like the conclusion we give above, candidates at this level do have some common sense to

Correlation of Features

Correlation is a measure of the linear relationship between 2 or more variables. Through correlation, we can predict one variable from the other. The logic behind using correlation for feature selection is that good variables correlate highly with the target. Furthermore, variables should be correlated with the target but uncorrelated among themselves.

If two variables are correlated, we can predict one from the other. Therefore, if two features are correlated, the model only needs one, as the second does not add additional information.



The above figure shows the Correlation coefficients between 6 features that we selected earlier. What we care about is how those features correlate to Score - which is what we are trying to predict in this problem. Unfortunately, though, the values are not good - albeit showing some correlations here. We believe this has something to do with the dataset - which does not represent the entire population. Therefore, we have decided not to use those features in our training.

2.5 Model selection and training

In this project, four common statistical classifiers are evaluated: Logistic Regression, Random Forests, K-NN and Linear SVC.

Logistic regression is a supervised classification model known as the logit model. It estimates the probability of something occurring, like 'will buy' or 'will not buy,' based on a dataset of independent variables. The outcome should be a categorical or a discrete value. K-NN algorithm is a basic yet essential classifier. It works by assuming the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available

categories.

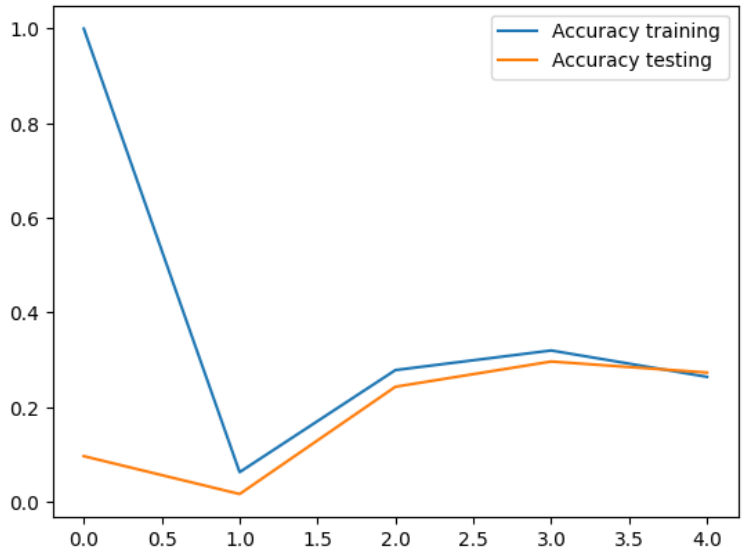
Random forests, one of the most popular ensemble classification techniques, combine a large number of decision-trees and bootstrap sampling to provide a low-bias, low-variance classification method. Regarding linear SVC, it is similar to SVC but with linear kernel, so it has more flexibility in the choice of penalties and loss functions and should scale better to large numbers of samples (in this case, however, the samples are not large). Like SVC, linear SVC can be extended to classify patterns that are not linear, which makes it suitable for the problem in question here.

In this study, the implementation of these classifiers was performed with the default settings. Due to the small dataset, stratified ten-fold cross validation method was used to evaluate the performance of each classifier. Stratified sampling is suitable for imbalanced dataset like the one this project is dealing with as class distribution are preserved through every split.

K-Nearest Neighbors

Due to the fact that this classifier is simple by nature (it is also called a *lazy learner algorithm*, it is important that we tune its parameter to take full advantage of the power of this algorithm. A key parameter to consider here is the number of k, which refers to the number of nearest neighbours to include in the majority of the voting process. The graph below show how the accuracy (based on the metrics from the SKLearn package) changes as k changes from 0 to 4:

It is obvious that k = 1 produces the worst results, with low training and testing accuracy. The metric reaches the maximum value at k = 3, then decreases with the increase of k. Based on this chart, we have decided to go with k = 3 for KNN.

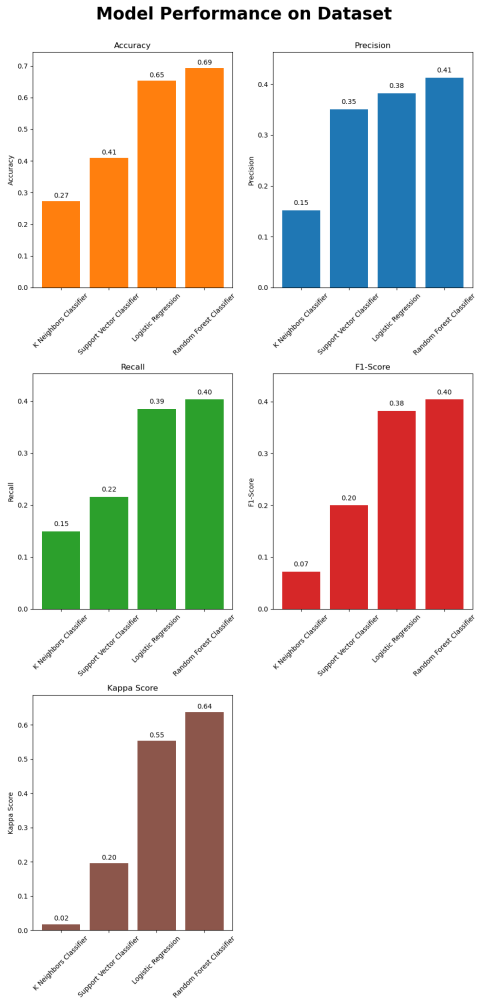


Chapter 3

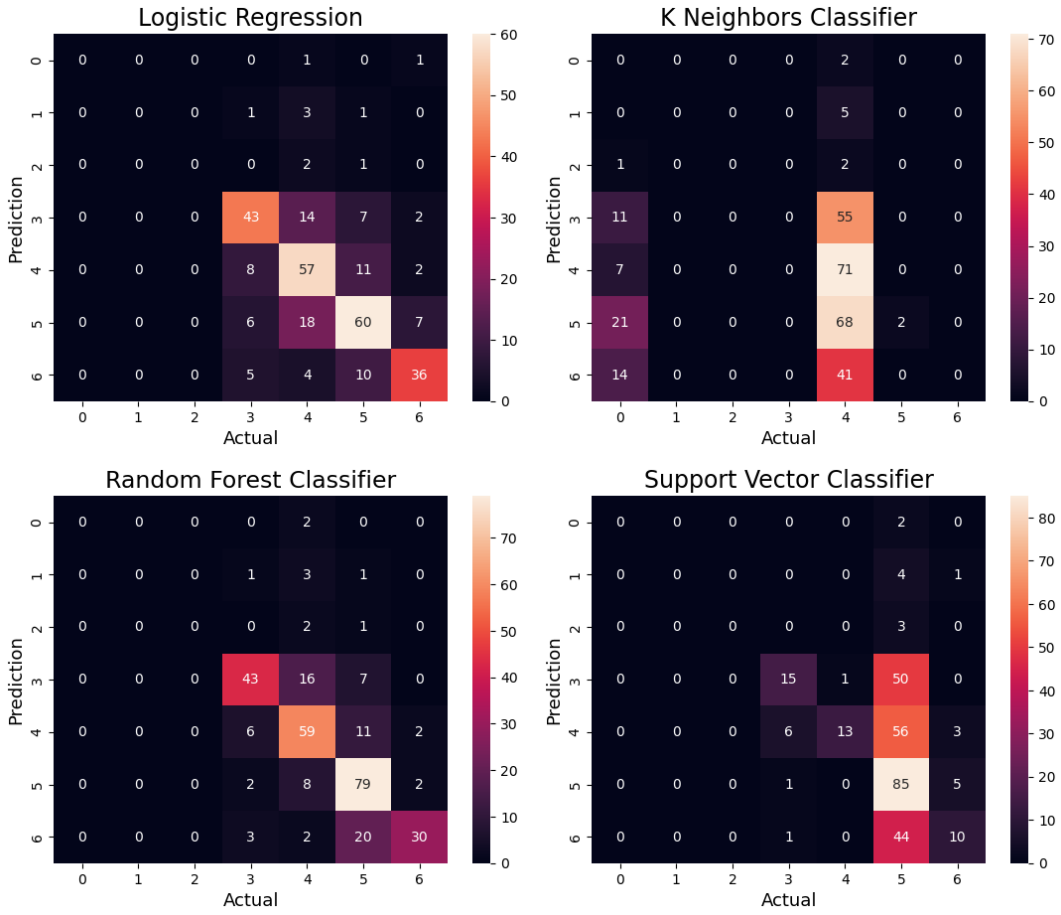
Results and further discussion

3.1 Evaluation results

The figure below shows the performance of 4 classifiers on test data.



The figure illustrates that the random forest classifier outperforms others, with accuracy and Kappa score reaching about 70%. Logistic Regression also produces similar results (Kappa score slightly worse than Random Forest). On the other hand, K Neighbours Classifier producing the worst performance, getting just roughly one fourth the number of predictions accurately. In other metrics, KNN are also placed last among 4 algorithms. All four classifiers in testing produce average precision, recall and F1-score. The confusion matrices below, each corresponds to one algorithm, corroborate the above statements:



In these matrices, the two Prediction and Actual axes represent normalized scoring from original values (0 corresponds to 4.0, 6 corresponds to 8.0). For Logistic Regression, we can see that it performs reasonably well on half of the score range. The same can be said for Random Forest Classifier. Meanwhile, the other two classifiers produce much worse results - it can only produce significant correct prediction on a single normalized value.

3.2 Further discussion and conclusion

Natural language processing is a hot topic in machine learning. It is also one of the most challenging one, as it has never been easy to teach machine something that is so fundemental of human kind. To some extent, this project has given a glimpse of what is possible with the help

of machine learning in a popular problem. The results are nowhere near excellent, which we do expect due to the nature of the dataset. In addition, this could be further improved should there be more time to explore further features and and tune the parameters. Nevertheless, its applications are promising. It is clear that automatic scoring is the future of language assessment, eliminating human bias out of the equation. In addition, this can be provided as a tool to support candidates with their preparation for IELTS test.

In the scope of this project, only Task 2 of the Writing component is analyzed in depth. In addition, some aspects of actual Writing assessment are not recreated. For example, the system do not detect plagiarism or topic relevance in the candidate's work. Future work should take this into account, as they are two important criteria in evaluating an IELTS writing response; academic writing in general. A larger and more inclusive dataset is necessary. Despite that, we argue that the evaluation is reasonable and give a good view of how various ML algorithms work in the context of this project.

Appendix A

Reference

[1] **New milestone for world-leading English test: IELTS trusted by more than 11,000 organisations across the globe**, IELTS, 2021, *https://www.ielts.org/news/2021/ielts-trusted-by-more-than-11000-organisations-across-the-globe*

[2] **English centers laugh all the way to the bank on booming IELTS demand**, VNExpress International, April 27, 2022, *e.vnexpress.net/news/trend/english-centers-laugh-all-the-way-to-the-bank-on-booming-ielts-demand-4453991.html*

[3] **IELTS Test format**, IELTS, 2023, *https://www.ielts.org/for-test-takers/test-format*

[4] **Sample test questions**, IELTS, 2023, *https://www.ielts.org/for-test-takers/sample-test-questions*

[5] **IELTS Writing key assessment criteria**, IELTS, 2023, *https://www.ielts.org/-/media/pdfs/ielts-writing-key-assessment-criteria.ashx*

[6] **IELTS Writing band descriptors**, IELTS, 2023, *https://www.ielts.org/-/media/pdfs/ielts-writing-band-descriptors.ashx*

[7] **IELTS Preparation**, Cambridge English, 2023, *https://www.cambridgeenglish.org/exams-and-tests/ielts/preparation/*

[8] **Test taker performance 2022**, IELTS, 2023, *https://www.ielts.org/for-researchers/test-statistics/test-taker-performance*

[9] Bishop, Christopher (2006). *Pattern recognition and machine learning.* Berlin: Springer. ISBN 0-387-31073-8.