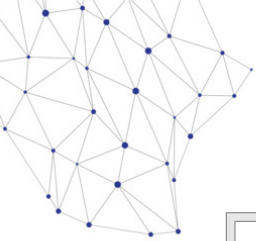# IELTS WRITING ASSESSMENT

Machine Learning Capstone Project

Group 9

Instructor: Prof. Khoat Than

September 23, 2023

- Problem Statement

- Exploratory data analysis

- Model Selection & Evaluation

- Discussion

Problem Statement

- IELTS is widely considered as the most popular English language test for migration and higher education, with huge demand in Vietnam

- Students find difficult to prepare for the Writing components due to multiple factors

Decision

Study a system that can assess candidates' Writing answers without human raters' presence.

- Problem: Automatically marking answers to IELTS Academic Writing tasks without the need for human raters.
- Problem type: Multiclass classification problem
- Input: An essay to assess
- Output: A corresponding band score

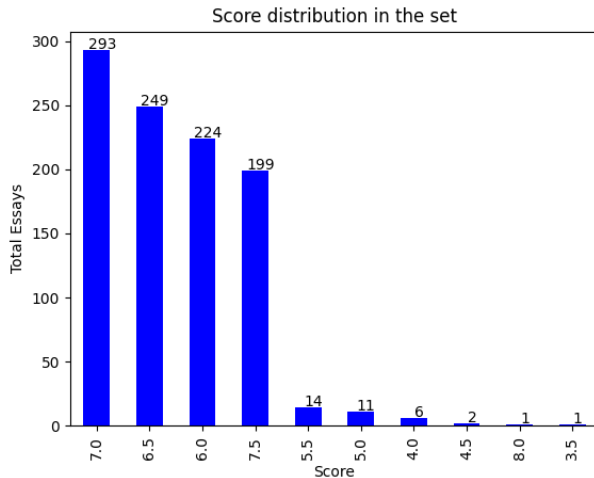- Source: Self-constructed from official preparation materials and websites

Minimal preprocessing:

- Converting to lowercase
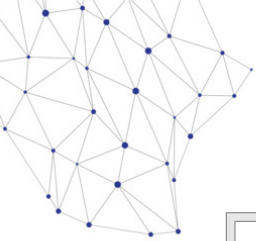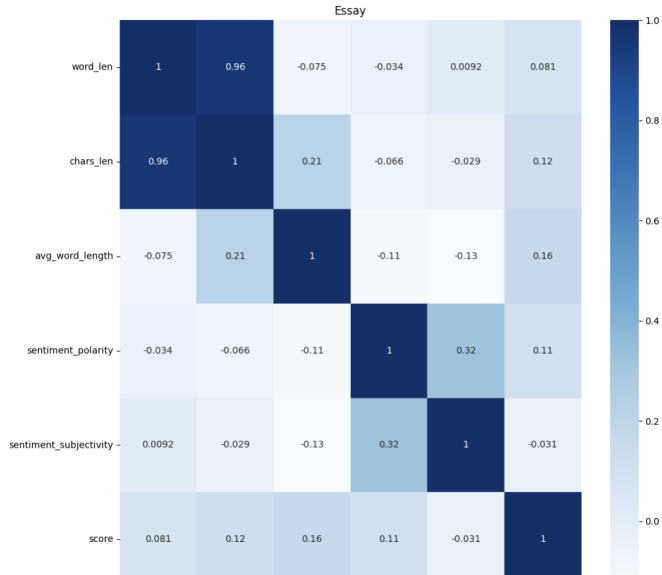- Using CountVectorizer in SKlearn as a tokenizer and feature extractor.

Exploratory data analysis

Score distribution

Score distribution in the set
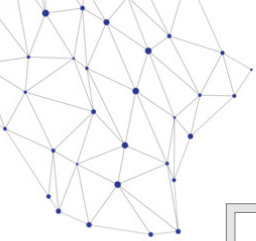
Feature Selection

Essay

Model Selection & Evaluation

- Estimates the probability of an event occurring, such as voted or didn't vote, based on a given dataset of independent variables.

- A non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point
- One of the most simple methods in ML.

- An ensemble method that take the average of multiple decision trees to make predictions. High-performance.
- A popular machine learning algorithm used for classification.

- A variant of SVM - a machine learning algorithm used both classification
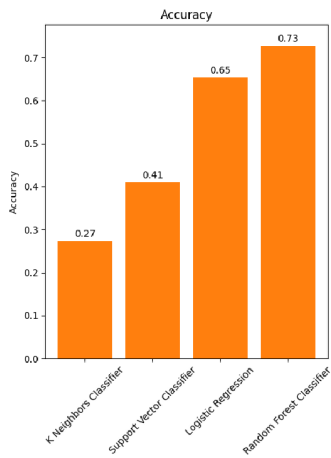- Popular for its ability to handle high-dimensional data
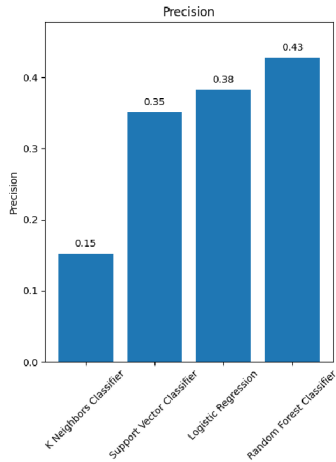- Based on the kernel linear method

Figure: Accuracy Performance

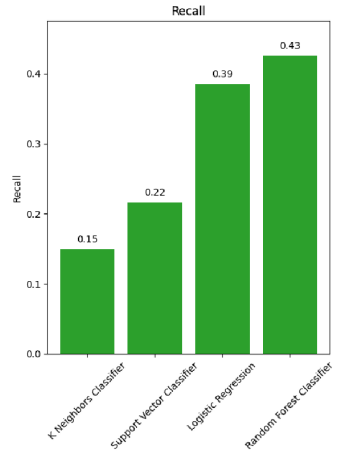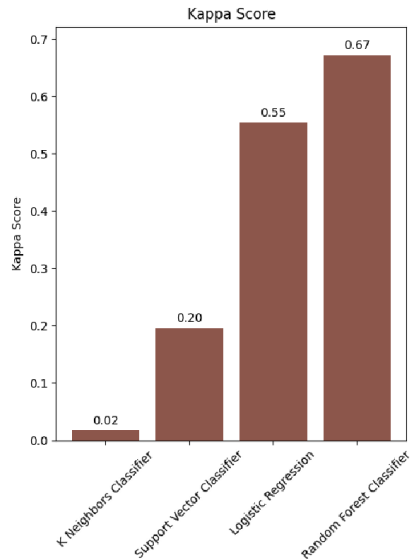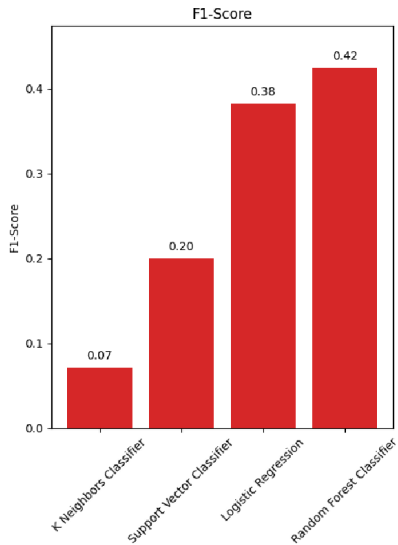
Figure: Precision Performance

Figure: Recall Performance

Model Selection & Evaluation

F1-Score

K Neighbors Classifier: 0.07
Support Vector Classifier: 0.20
Logistic Regression: 0.38
Random Forest Classifier: 0.42

Kappa Score

K Neighbors Classifier: 0.02
Support Vector Classifier: 0.20
Logistic Regression: 0.55
Random Forest Classifier: 0.67
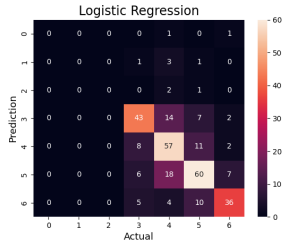
Model Selection & Evaluation
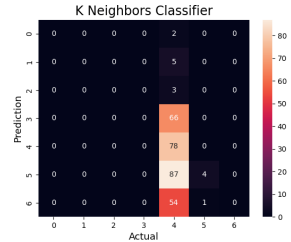
Figure: Confusion Matrix Logistic Regression

Figure: Confusion Matrix K Nearest Neighbors
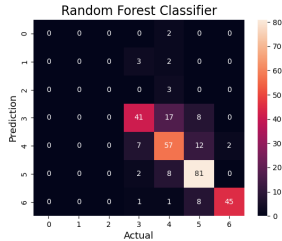
Model Selection & Evaluation

Figure: Confusion Matrix
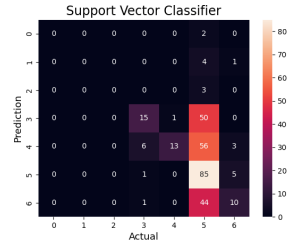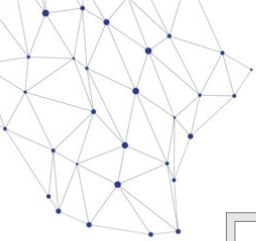Random Forest Classifier

Figure: Confusion Matrix
Support Vector Classifier

Conclusion

- To some extent, this project has given a glimpse of what is possible with the help of machine learning in a popular problem.
- A lot of room for improvement in terms of performance
- Its applications are still promising

THANKS FOR
YOUR ATTENTION

Dang Manh Cuong 20214949

Trinh Truong Giang 20214958

Tran Minh Tuan 20214978