|  |  |  |  |
| --- | --- | --- | --- |
| LOGO DHCNTT -hinh.jpg | ĐẠI HỌC QUỐC GIA TP. HCM  **TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN** | Ngày nhận hồ sơ |  |
| *(Do CQ quản lý ghi)* | |

**THUYẾT MINH**

ĐỀ TÀI KHOA HỌC VÀ CÔNG NGHỆ CẤP SINH VIÊN 2022

# THÔNG TIN CHUNG

## A1. Tên đề tài

* Tên tiếng Việt (IN HOA): PHÂN TÍCH SỰ ẢNH HƯỞNG CÁC ĐẶC TRƯNG TRÍCH XUẤT TỪ AUDIO LUYỆN NÓI TIẾNG ANH VỚI ĐIỂM BÀI THI NÓI IELTS.
* Tên tiếng Anh (IN HOA): ANALYZE THE INFLUENCE OF FEATURES EXTRACTED FROM AUDIOS ON THE RESULTS OF IELTS SPEAKING TEST.

## A2. Thời gian thực hiện

**..06..** tháng (kể từ khi được duyệt).

## A3. Tổng kinh phí

*(Lưu ý tính nhất quán giữa mục này và mục B8. Tổng hợp kinh phí đề nghị cấp)*

Tổng kinh phí: …**6**.. triệu đồng,gồm

* Kinh phí từ Trường Đại học Công nghệ Thông tin: ..**6**.. triệu đồng

## A4. Chủ nhiệm

Họ và tên: **Nguyễn Ngọc Hà My** **.**

Ngày, tháng, năm sinh: 08-03-2002 . Giới tính (Nam/Nữ): Nữ .

Số CCCD: 051302005048 ; Ngày cấp: 12-08-2021 ; Nơi cấp: Cục trưởng cục cảnh sát quản lý hành chính về trật tự xã hội .

Mã số sinh viên: 20521623 **.**

Số điện thoại liên lạc: 0828924893 **.**

Đơn vị (Khoa): KH&KTTT **.**

Số tài khoản: 1019521800 Ngân hàng: Vietcombank

## A5. Thành viên đề tài

|  |  |  |  |
| --- | --- | --- | --- |
| **TT** | **Họ tên** | **MSSV** | **Khoa** |
| 1 | Nguyễn Thị Phương Thảo | 20521936 | KH&KTTT |
| 2 | Trần Thị Thu Hà | 20521273 | KH&KTTT |

# 

# MÔ TẢ NGHIÊN CỨU

## B1. Giới thiệu về đề tài

*(Ghi các ý về tổng quan tình hình nghiên cứu liên quan đến đề tài, lí do thực hiện đề tài, các thách thức)*

### B1.1. Tổng quan tình hình nghiên cứu

Trong thời đại 4.0 ngày nay, nhu cầu sử dụng thành tạo tiếng Anh trong giao tiếp, học tập, làm việc,… ngày càng tăng kéo theo đó là sự phổ biến các chứng chỉ tiếng Anh quốc tế như IELTS, TOEIC, TOEFL cũng ngày càng phổ biến trong việc kiểm tra trình độ tiếng Anh của người làm việc tại nước ngoài, tuyển sinh đại học và sau đại học,… Tuy nhiên, chi phí để dự thi các kỳ thi này không hề rẻ nên người học thường phải trải qua quá trình ôn luyện kỹ càng trước khi bước vào một kỳ thi thực tế mà kỹ năng nói lại là một kỹ năng đòi hỏi người học cần sự tương tác và nhận xét từ giảng viên để có thể cải thiện hơn khả năng của mình. Trong các trung tâm luyện tiếng Anh, số lượng giáo viên có hạn nhưng hằng ngày phải nghe và nhận xét từ hàng chục đến hàng trăm audio của học sinh mỗi ngày khiến công việc này trở nên nhàm chán và áp lực. Nhận thấy điều đó, nhóm chúng tôi đã đề xuất việc chấm điểm thi nói IELTS tự động bằng việc trích xuất các đặc trưng có thể thu được từ audio và sử dụng mô hình máy học để dự đoán điểm thi một cách chính xác nhất.

### B1.2. Lý do thực hiện đề tài

Việc xây dựng một mô hình có thể dự đoán điểm thi IELTS speaking không những có thể giảm được chi phí luyện thi IELTS cho học sinh, đặc biệt là những học sinh tự ôn luyện tại nhà mà còn có thể giảm bớt áp lực cho các giáo viên khi họ có nhiều thời gian hơn để cải tiến phương pháp luyện thi và nâng cao chất lượng giảng dạy. Tuy nhiên, hiện nay dữ liệu thi IELTS speaking từ Hội đồng Anh không được công bố ra bên ngoài nên để xây dựng mô hình dự đoán này thì trước tiên cần xây dựng bộ dữ liệu về điểm thi IELTS speaking. Dữ liệu thô chúng tôi có là audio luyện thi IELTS từ trung tâm luyện thi, với số điểm là của học viên sau khi học và thi thật báo cáo về trung tâm hoặc là điểm khi học viên tham gia thi thử và được các giáo viên trung tâm chấm. Từ dữ liệu này, chúng tôi sử dụng các mô hình và các API để trích xuất ra các đặc trưng mong muốn từ dữ liệu là âm thanh nhằm xây dựng một bộ dữ liệu phục vụ cho mục đích tạo ra một mô hình có thể dự đoán điểm thi IELTS từ audio.

### B1.3. Các thách thức

**Về dữ liệu:**

* Các audio này do các học viên tự thu âm cho nên chất lượng nhiều audio không ổn định, audio có lẫn tạp âm, … làm ảnh hưởng đến kết quả.
* Sự thiếu hụt dữ liệu cần thiết để máy có thể học được và dự đoán chính xác. Dữ liệu về học viên có trình độ nằm trong khoảng 5.0-7.0 rất nhiều trong khi đó lại thiếu hụt trầm trọng các nhãn 0-3.0 hay nhãn 8.0-9.0 khiến cho mô hình chỉ dự đoán tốt đối với các trường hợp 5.0-7.0.

**Về phương pháp:** Tìm cách chuyển từ audio sang text. Quá trình này không hề đơn giản khi mà tiêu chí chấm điểm thi lại yêu cầu đến yếu tố ngữ nghĩa như độ mạch lạc của bài nói; bài nói có liên quan đến câu hỏi không; nhận diện và đánh giá mức độ khó của các cấu trúc câu, từ, ngữ pháp trong bài khi trong tiếng Anh những cấu trúc ngữ pháp tuy là một nhưng có sự khác nhau về động từ, danh từ, thì… Những bài toán trên vẫn còn là những vấn đề nan giải trong xử lý ngôn ngữ tự nhiên. Cho nên vấn đề này yêu cầu cần giải quyết nhiều bài toán nhỏ phức tạp để có thể trích xuất ra được đầy đủ các tiêu chí cần thiết cho việc đánh giá điểm thi một cách chính xác.

**Về ứng dụng thực tiễn:** Hiện tại, đã có tồn tại một số công cụ trên web và trên di động không kiểm chứng được độ chính xác, không công bố cách chấm. Cần phải làm một nghiên cứu cụ thể kiểm chứng độ chính xác.

## B2. Mục tiêu, nội dung, kế hoạch nghiên cứu

### B2.1 Mục tiêu

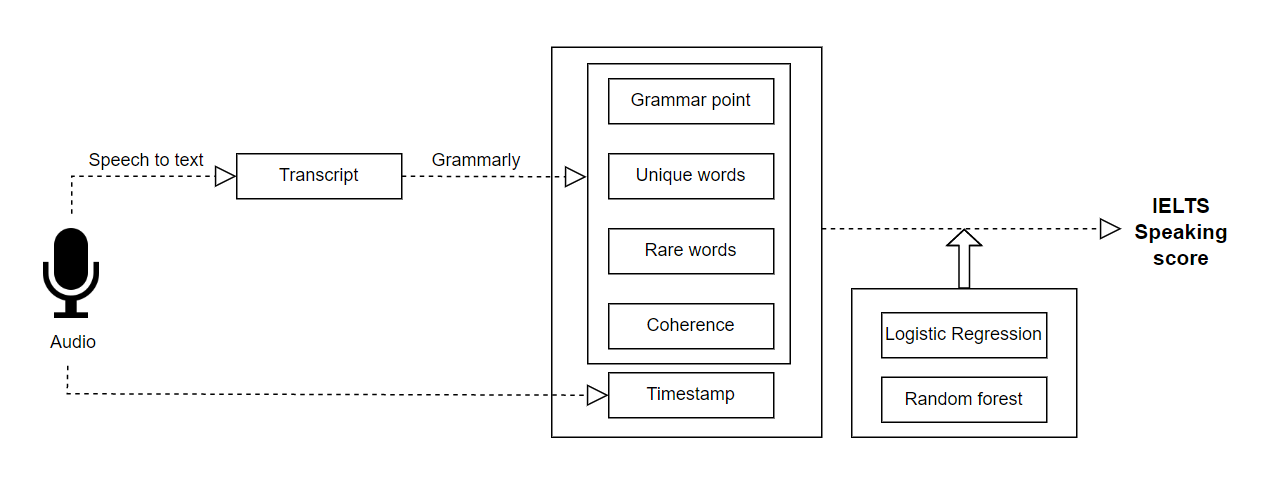
Trong phạm vi đề tài này, chúng tôi sẽ:

* Xây dựng một bộ dữ liệu bao gồm thông số các đặc trưng trích xuất được từ audio luyện nói tiếng Anh (thông tin và âm thanh của học viên được bảo mật).
* Đề xuất các phương pháp trích xuất một số đặc trưng từ audio nhằm đánh giá kỹ năng nói.
* Phân tích ảnh hưởng của các đặc trưng trích xuất được đến điểm thi IELTS.
* Đề xuất các phương pháp để xử lý dữ liệu không cân bằng phù hợp với mô hình dự đoán cho bộ dự liệu.
* Đề xuất mô hình máy học sử dụng để dự đoán điểm thi IELTS trên bộ dữ liệu này.
* Thực nghiệm, đánh giá và phân tích kết quả.

### B2.2 Nội dung và phương pháp nghiên cứu

**Nội dung 1:** Xây dựng bộ dữ liệu bằng cách trích xuất các đặc trưng từ audio

Dữ liệu thô chúng tôi có được là các audio luyện nói của học viên và điểm thi IELTS của họ trong kỳ kiểm tra ở trường và được các giảng viên có chuyên môn ở trung tâm chấm điểm, ngoài ra còn một số điểm dữ liệu và điểm thi thật của học viên sau khi hoàn thành khóa học. Dựa trên các tiêu chí đánh giá bài thi nói IELTS mà hội đồng Anh công bố, chúng tôi nhận thấy có thể trích xuất vài đặc trưng dựa trên các công cụ có sẵn để đánh giá các tiêu chí như độ lưu loát, từ vựng, độ chính xác ngữ pháp. Quá trình thực hiện được mô tả như hình dưới đây:



Hình 1: Sơ đồ xây dựng dữ liệu và dự đoán điểm thi.

**Phương pháp:**

1. **Trích text:**

Quá trình trích xuất ra bản chép lời từ dữ liệu âm thanh gọi là speech-to-text. Bản chép lời từ các đoạn audio IELTS speaking là một trong những thuộc tính vô cùng quan trọng phục vụ cho việc đánh giá tiêu chí coherence, lexical resource, grammartical range and accuracy vì các yếu tố này khi được đánh giá trên đoạn chép lời (có qua tiền xử lý) sẽ thuận tiện và chính xác hơn nhiều.

Để trích xuất bản chép lời từ audio, chúng tôi sử dụng thư viện SpeechRecognition có sẵn trên Python.

* Chuyển audio từ các dạng .mp3, .mp4, m4a sang dạng .wav.
* Sử dụng hàm recognize\_google để trích text.

**Tiền xử lý dữ liệu:** Bản chép lời sau khi được trích xuất vẫn còn nhiều điểm cần tiền xử lý như: phát âm sai dẫn đến trích xuất từ chưa đúng, chất lượng âm thanh kém nên có đoạn bị mất text, thiếu dấu câu như chấm (.), phẩy (,), gạch nối (-), học viên nói bị lặp dẫn đến text cũng bị lặp, audio có xen lẫn phần câu hỏi của giám khảo. Các yếu tố này sẽ ảnh hưởng đến việc đánh giá bài thi nói dựa trên yếu tố ngữ nghĩa nên cần qua bước tiền xử lý rồi mới tiến hành trích xuất dữ liệu về ngữ nghĩa.

Công việc của tiền xử lý bao gồm thêm dấu câu, loại bỏ từ bị lặp, loại bỏ text của câu hỏi, thêm text bị thiếu. Bản chép lời sau khi tiền xử lý sẽ được tiếp tục kiểm tra lỗi ngữ pháp và từ vựng.

1. **Lexical resource:**

Yếu tố lexical resource trong tiêu chí đánh giá điểm thi IELTS tập trung vào đánh giá từ vựng được sử dụng trong bài nói có phong phú không? Có bao gồm nhiều từ khó, từ ít thông dụng không?

Để giải quyết các câu hỏi trên, chúng tôi sử dụng Bag of words để tính toán thuộc tính %unique words ( % số từ khác nhau được sử dụng) và dựa trên bộ dữ liệu về các từ vựng bậc cao trong tiếng Anh để tính toán % rare words (% số từ khó được sử dụng).

1. **Độ mạch lạc của bài nói:**

* Tìm hiểu các mô hình trong NLP, deep learning nhằm chấm điểm mạch lạc của bài nói.
* Xây dựng bộ dữ liệu có gán nhãn các câu có liên quan với nhau và các câu không liên quan với nhau
* Chấm điểm dựa trên các tiêu chí như: có sử dụng từ nối hay không? các câu trong đoạn có cùng chủ đề hay không? Có trả lời được câu hỏi hay không?
* Sử dụng phương pháp Machine learning, Deep learning và đánh giá độ phù hợp của các mô hình học máy, học sâu này đối với việc chấm điểm mạch lạc của bài nói.

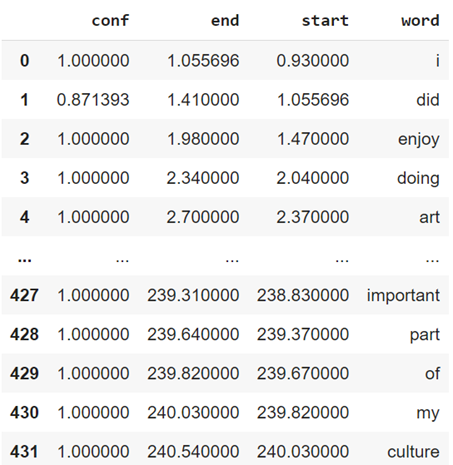
1. **Độ lưu loát:**

**B1**: Trích timestamp

Sử dụng thư viện vosk trên Python trích xuất ra các yếu tố:

* Từ (word)
* Thời gian bắt đầu (start)
* Thời gian kết thúc (end)
* Độ dứt khoát (conf)

Kết quả minh họa:



Hình 2: Tổng quan về kết quả các đặc trưng trích xuất đặc bằng thư viện vosk

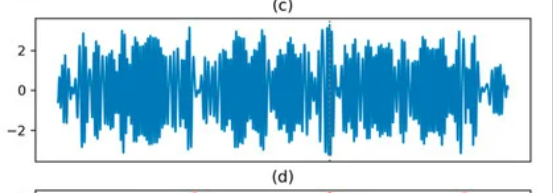
**B2**: Tính toán độ ngập ngừng

Từ dữ liệu đã trích được, ta có thể tính độ ngập ngừng giữa các từ và từ đó đánh giá tiêu chí fluency của bài nói. Độ ngập ngừng ở đây được tính bằng công thức start từ đứng sau – end từ đứng trước, ta thu được kết quả là thuộc tính độ ngập ngừng. Độ ngập ngừng càng cao chứng tỏ độ lưu loát thấp và ngược lại.

1. **Sử dụng nối âm:**

Việc sử dụng nối âm từ âm đuôi của từ trước nối với âm đầu của từ sau là một trong những thủ thuật cơ bản mà những người nói tiếng Anh thành thạo hay dùng nhằm tăng tốc độ nói cũng như cải thiện độ ngập ngừng. Việc đánh giá khả năng nối âm của học viên cũng là một trong những yếu tố có thể dùng để đánh giá khả năng phát âm và độ trôi chảy của bài nói. Để xem xét tự động người nói có nối âm hay không chúng tôi đề xuất:

* Thu thập dữ liệu âm thanh về các trường hợp có thể nối âm với nhau gồm âm thanh có sử dụng nối âm và âm thanh không sử dụng nối âm.
* Phân tích phổ (spectrum) của dữ liệu âm thanh trên để xây dựng mô hình dự đoán đoạn phổ nào có chứa âm thanh sử dụng nối âm.
* Từ text của bài nói trích xuất được như đã trình bày ở mục 1., xem xét phổ của các trường hợp có thể sử dụng nối âm và áp dụng mô hình dự đoán trên để tính tần suất sử dụng nối âm của người nói.



Hình 3: phân tích phổ của đoạn âm thanh

Kết quả dự kiến: Bộ dữ liệu dùng để dự đoán điểm thi IELTS với input (x) là các đặc trưng nêu trên.

**Nội dung 2**: Phân tích ảnh hưởng của các đặc trưng đến điểm thi IELTS

Phương pháp: Sử dụng ANOVA for regression và lack of fit để kiểm định độ phù hợp của mô hình hồi quy

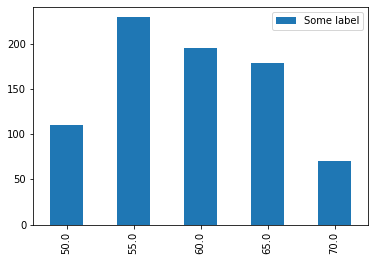
* Phân tích phương sai (ANOVA) bao gồm các phép tính cung cấp thông tin về mức độ biến thiên trong mô hình hồi quy và tạo cơ sở cho các kiểm định về mức độ ý nghĩa

Kết quả dự kiến:

* Bảng kết quả về kiểm định độ tương quan giữa từng đặc trưng với điểm thi IELTS.
* Bảng kết quả về kiểm định độ tương quan giữa các đặc trưng với nhau.

**Nội dung 3**: Chọn ra kỹ thuật, phương pháp xử lý dữ liệu không cân bằng phù hợp nhất cho bộ dữ liệu

Dữ liệu của chúng tôi gặp trường hợp bị mất cân bằng biến mục tiêu (Xem Hình 4). Dữ liệu ghi lại của các học viên có trình độ nói tiếng Anh thấp (từ 0.0 đến 4.5) và trình độ nói tiếng Anh lưu loát (từ 7.5 đến 9.0) ít được ghi nhận từ nguồn mà chúng tôi thu thập. Dẫn đến hiện tượng các mô hình dự đoán chỉ tập trung Các giải thuật Machine Learning và Deep Learning thường sẽ có xu hướng cải thiện độ chính xác bằng cách giảm độ lỗi mà không quan tâm đến phân phối của biến mục tiêu. Điều đó dẫn đến hiện tượng các kỹ thuật học máy thiên về dự đoán biến mục tiêu có giá trị quan sát được là đa số và sẽ gặp dự đoán sai ở hầu hết các điểm dữ liệu có biến mục tiêu đạt giá trị quan sát được là thiểu số.



Hình 4: Biểu đồ phân bố số lượng nhãn.

Phương pháp: Các phương pháp có thể áp dụng đối với dữ liệu không cân bằng có thể là: tiền xử lý dữ liệu không cân bằng, điều chỉnh thuật toán, hậu xử lý dữ liệu. Các phương pháp phổ biến hiện nay đều dựa trên tiền xử lý dữ liệu không cân bằng. Trong phần này, chúng tôi sẽ áp dụng các phương pháp tiền xử lý- resampling lại dữ liệu như SMOTE, Near Miss và thuật toán LDS để training dữ liệu không cân bằng, sau đó chạy thực nghiệm và so sánh kết quả.

1. SMOTE (Synthetic Minority Oversampling Technique) là một trong những phương pháp phổ biến được sử dụng để giải quyết vấn đề mất cân bằng. Nó tăng ngẫu nhiên các mẫu thiểu số bằng cách sao chép chúng dựa trên k- hàng xóm gần nhất (k\_nearest neighbors) cho mỗi ví dụ của lớp thiểu số.
2. Near Miss là một trong những phương pháp cân bằng phân phối dữ liệu bằng cách giảm số lượng mẫu. Nó loại bỏ ngẫu nhiên các mẫu có giá trị mục tiêu quan sát được là đa số dựa trên khoảng cách giữa các lớp khác nhau. Khi các quan sát được của hai lớp khác nhau rất gần nhau, nó loại bỏ các quan sát của lớp đa số để tăng khoảng cách giữa hai lớp.
3. Thuật toán LDS được công bố bởi Yuzhe Yang và cộng sự, sử dụng ước tính mật độ kernel để tìm hiểu sự ảnh hưởng mất cân bằng trong các bộ dữ liệu tương ứng với các biến mục tiêu liên tục. LDS kết hợp sử dụng một kernel đối xứng với phân phối mật độ quan sát được để trích xuất một phiên bản được làm mịn giải thích cho sự chồng chéo thông tin của các mẫu dữ liệu của nhãn lân cận. Kernel đối xứng là bất kỳ kernel nào thỏa mãn:

Sau đó với mỗi mẫu ( thuộc tập train, tính toán trọng số weight cho mỗi mẫu

Những mẫu nào có biến mục tiêu thuộc giá trị xuất hiện càng phổ biến thì sẽ có trọng số w này càng thấp và ngược lại, những mẫu nào có biến mục tiêu thuộc giá trị xuất hiện hiếm thì sẽ có trọng số w cao. Trong quá trình training, sử dụng hàm loss với trọng số weight này

Kết quả dự kiến:

* Bảng kết quả khi áp dụng các phương pháp trên cho các mô hình khác nhau.
* Biểu đồ độ lỗi của mô hình dự đoán với từng nhãn khi áp dụng các phương pháp trên.
* Tìm ra được phương pháp thích hợp nhất để xử lý vấn đề mất cân bằng của bộ dự liệu.

**Nội dung 4**: Đề xuất mô hình máy học để dự đoán điểm thi IELTS

Phương pháp: Chia bộ dữ liệu thành tập train và tập test chứa số lượng các giá trị output (y) phù hợp. Các mô hình chúng tôi đề xuất sử dụng trong mô hình này bao gồm:

* Linear regression: Hồi quy tuyến tính là một mô hình tuyến tính, ví dụ: một mô hình giả định mối quan hệ tuyến tính giữa các biến đầu vào (x) và biến đầu ra duy nhất (y). Cụ thể hơn, y có thể được tính toán từ sự kết hợp tuyến tính của các biến đầu vào (x). Trong khi dữ liệu chúng tôi có là thông số các đặc trưng tự trích xuất được và kết quả dự đoán là điểm thi, việc áp dụng mô hình hồi quy là một trong những mô hình khả thi để giải quyết bài toán
* Logistic Regression for multi-class classification: Tuy đầu ra (y) cần dự đoán là điểm (giá trị số), giá trị của (y)- điểm thi IELTS chỉ nằm trong khung từ 0-9.0 và không liên tục (chỉ có các điểm như 9.0, 8.5, 8.0, …, 0) cho nên có thể cho output (y) là nhãn chứa điểm và áp dụng mô hình phân lớp để giải quyết bài toán.
* Random forest: mô hình rừng ngẫu nhiên cũng là một mô hình dùng để giải quyết bài toán phân lớp. Trong quá trình thu thập và khảo sát dữ liệu, chúng tôi nhận thấy rằng tuy khác lớp (khác output (y)), giá trị của một số đặc trưng có thể không thay đổi nhiều. Điều này có thể gây khó khăn cho việc dự đoán khi sử dụng 2 mô hình trên nên đã đề xuất việc sử dụng thêm mô hình rừng ngẫu nhiên và nhận thấy kết quả có khả quan hơn.

Sử dụng độ đo RMSE và R square để đánh giá độ phù hợp của mô hình đối với dữ liệu.

Kết quả dự kiến:

* Bảng kết quả dự đoán của các mô hình trên độ đo RMSE và R square.
* Ma trận nhầm lẫn của mô hình Random forest và Logistic regression trên tập test.
* Đánh giá về độ phù hợp của các mô hình và phân tích nguyên nhân, kết quả.
* Đề xuất ra những phải pháp đối với bộ dữ liệu (undersampling, oversampling) để cải thiện kết quả.

**B2.3 Kế hoạch nghiên cứu**.

|  |  |
| --- | --- |
| **Thời gian** | **Nội dung thực hiện** |
| **Giai đoạn 1:**  Tuần 1 – Tuần 4: | * Thu thập audio. * Tìm hiểu cách trích text từ audio. * Tìm hiểu các đặc trưng của audio. |
| **Giai đoạn 2:**  Tuần 5 – Tuần 9: | * Tìm hiểu cách sử dụng của thư viện vosk để trích xuất từng đặc trưng. * Phân tích từng đặc trưng và mối liên hệ của các đặc trưng với nhau bằng phương sai ANOVA. |
| **Giai đoạn 3:**  Tuần 10 – Tuần 15: | * Tìm hiểu cái mô hình đơn giản để kiểm thử bộ dữ liệu như: Linear regression, Logistic Regression for multi-class classification, Random forest. * Tìm hiểu các độ đo: RMSE, R square và thực nghiệm kết quả theo thuộc tính của dữ liệu. |
| **Giai đoạn 4:**  Tuần 16 – Tuần 20: | * Đánh giá và phân tích kết quả thực nghiệm. * Chỉnh sửa thuộc tính để bộ dữ liệu có kết quả chạy tốt nhất. |
| **Giai đoạn 5:**  Tuần 21 – Tuần 24: | * Đưa ra hướng phát triển cho bộ dữ liệu. * Kết luận. * Viết báo cáo. |

## B3. Kết quả dự kiến

* Bộ dữ liệu về các đặc trưng trích xuất được và điểm thi nói IELTS.
* Bảng kết quả dự đoán của các mô hình Logistic regression, Linear regression, Random forest trên độ đo RMSE và R square.
* Kết luận về ảnh hưởng của các đặc trưng trích xuất được với điểm thi nói IELTS.
* Phân tích và đánh giá độ phù hợp của các mô hình máy học đối với bộ dữ liệu.

## B4. Tài liệu tham khảo

[1] Shadiev, Rustam, et al. "Review of speech-to-text recognition technology for enhancing learning." Journal of Educational Technology & Society 17.4 (2014): 65-84.

[2] O. Lyashevskaya, I. Panteleeva, and O. Vinogradova, “Automated assessment of learner text complexity,” Assessing Writing, vol. 49, Article ID 100529, 2021.

[3] S. C. Weigle, “English language learners and automated scoring of essays: c,” Assessing Writing, vol. 18, no. 1, pp. 85–99, 2013.

[4] P. Bamdev, M. S. Grover, Y. K. Singla, P. Vafaee, M. Hama, and R. R. Shah, “Automated speech scoring system under the lens: evaluating and interpreting the linguistic cues for language proficiency,” 2021.

[5] K. K. Y. Chan, T. Bond, and Z. Yan, “Application of an automated essay scoring engine to English writing assessment using many-facet rasch measurement,” Language Testing, 2022.

|  |  |
| --- | --- |
| *Ngày 22 tháng 09 năm 2022*  **Giảng viên hướng dẫn**  (Ký và ghi rõ họ tên) | *Ngày 22 tháng 09 năm 2022*  **Chủ nhiệm đề tài**  (Ký và ghi rõ họ tên) |
| Đỗ Trọng Hợp | Nguyễn Ngọc Hà My |
|  |  |