Project Objective

Develop a machine learning system that scores grammar quality from spoken English audio clips. Two distinct approaches were explored and optimized:

FIRST MODEL: AUDIO-FEATURE-BASED REGRESSION

1. Approach

implemented a regression pipeline leveraging **handcrafted acoustic features** directly from audio files.

Highlights:

- Extracted **40+ features** using librosa, including:
  - **MFCCs** (20 coefficients × mean + std = 40)
  - Spectral centroid, bandwidth, roll-off
  - Zero-crossing rate (ZCR), RMSE, tempo, and duration
- Imputed missing values using column-wise means.
- Selected **top 30 features** via SelectKBest using **mutual information**.
- Trained a **stacked ensemble** combining:
  - SVR, GradientBoostingRegressor, and HistGradientBoostingRegressor
  - Final estimator: RidgeCV

2. Preprocessing Steps

- **Audio Loading**: Resampled to 16kHz mono for uniformity.
- **Feature Extraction**: Hand-engineered from waveform using librosa.
- **Missing Values**: Filled NaNs with feature-wise means.
- **Feature Selection**: Used mutual_info_regression to retain 30 informative features.

3. Model Optimization

- Hyperparameter tuning via RandomizedSearchCV on:
  - learning_rate, max_depth, min_samples_leaf, l2_regularization
- Evaluation via **5-fold cross-validation**.

4. Feature Importance

- Top contributors:
  - MFCCs (especially lower orders)
  - Spectral centroid
  - ZCR and RMSE
- **Permutation importance** used to quantify feature influence.

Summary

- Strong performance using classical ML without deep learning.

- Interpretability preserved with handcrafted features.
- Modular and reproducible pipeline using scikit-learn.

SECOND MODEL: TRANSCRIPTION + NLP FEATURES

1. Approach

This approach leverages **speech-to-text** conversion followed by **linguistic feature analysis**.

Steps:

1. **Transcription** using a pretrained ASR model (model.transcribe(...))
2. **Feature Extraction** from transcript using nltk:
   - Total word count
   - Number of nouns, verbs, adjectives
   - Average word length
3. **Prediction** using a RandomForestRegressor.

2. Preprocessing

Training Phase:

- Transcribed audio files and extracted linguistic features.
- Removed raw text post-feature extraction.
- Handled missing values via mean imputation.
- Applied StandardScaler.

.

3. Model and Metrics

- **Model**: RandomForestRegressor
- **Hyperparameters**:
  - n_estimators=200, max_depth=10, random_state=42
  -

4. Training Evaluation:

Metric	Value
RMSE	0.5030
MAE	0.4144
R^2	0.7695
MAPE	13.15%

5. Feature Importance

Feature	Importance
total_words	High
num_verbs	High
avg_word_len	Moderate
num_nouns	Moderate
num_adjs	Low

Insight: Grammar scores correlated most with **word quantity** and **verb usage**.

Optimization Takeaways

Model	Strength	Weakness
Model 1 (Audio features)	High interpretability, no reliance on ASR	Sensitive to noise; indirect grammar cues
Model 2 (Transcription + NLP)	Direct grammar relevance, better contextual modeling	Dependent on transcription quality

Improvement Journey:

- Started with signal-based features (Model 1).
- Shifted toward linguistically interpretable features (Model 2).
- Achieved better alignment with grammar scoring task through NLP.

Optimization Efforts Summary

Initially, I built a grammar scoring model based purely on **handcrafted audio features** extracted from speech samples. This included MFCCs, spectral properties, and rhythm-based statistics. I trained a **stacked ensemble regressor** (SVR, Gradient Boosting, Ridge) and achieved modest predictive performance, but found the model lacked direct insight into grammatical structure.

To improve accuracy, I transitioned to a second model that transcribed the audio into text using an **automatic speech recognition (ASR)** system. From the transcriptions, I extracted **linguistic features** such as word counts, part-of-speech statistics (e.g., number of verbs and nouns), and average word length. I then trained a **RandomForestRegressor** on these features. This approach directly targeted grammar-related signals and yielded a significant improvement in accuracy. The RMSE dropped to **0.503**, MAE to **0.414**, and the model achieved an R^2 of **0.769**, a strong indicator of better generalization. Feature importance analysis also showed that verb and noun frequency were strong predictors of grammar quality.

By shifting from indirect audio-based proxies to direct linguistic analysis, I was able to optimize the grammar score predictions more effectively and align the model with human-like evaluation.