Dự Đoán Thể Loại Nhạc Bằng Machine Learning

- Nguyễn Tuấn Đạt, Châu Hải Đăng, Trần Đại Thắng, Nguyễn Xuân Đăng -

Problem Review

Thuộc dạng bài toán hồi quy (Classification)

Input:

- Dữ liệu nhà cửa và khu vực xung quanh gồm thông tin cá nhân từ **Kaggle Music Genre Classification.**

Output:

- Dự đoán giá nhà.

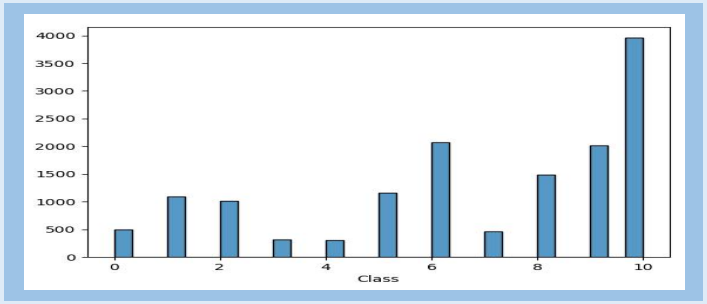
Notice:

- Tập dữ liệu bị thiếu, mất mát và nhiễu. Cần tiền xử lí và phân tích đặc trưng.

Dataset

Source:

- Kaggle – Musical Genre Classification.
https://www.kaggle.com/competitions/shai-music-genre-classification

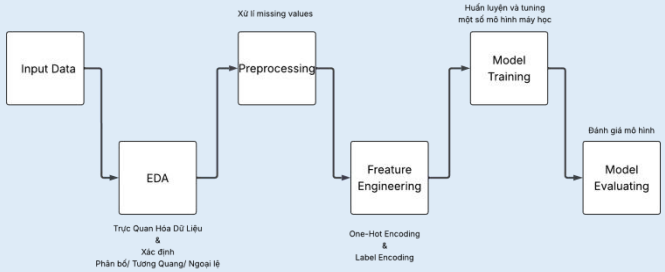


Mô tả:

- Gồm 14396 bộ train và 3600 bộ test.
- Bao gồm dữ liệu liên tục/ rời rạc và Chuỗi

Loại Dữ Liệu	Số lượng
Số	16
Chuỗi	2

Propose Method



Exploring Data Analysis

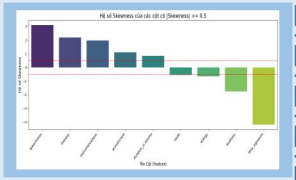
Inspecting Data:

Cần Chú Ý	Đặc trưng
Dữ liệu có skewness cao	- Lệch phải: duration, liveness, speechinessm acousticness, instrumentalness, - Lệch trái: Loudness, Energy
Phân phối chuẩn	- Popularity, danceability, valence, tempo
Dữ liệu rời rạc	- time_signature, mode, key.

Missing Data:

Đặc trưng	Train	Test
popularity	333	95
key	1609	405
instrumentalness	3541	836

Skewness:

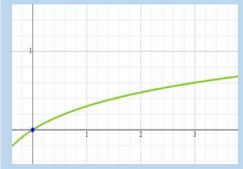


Preprocessing

Missing Datas:

Popularity, Key, Instrumentalness
- Fill median.

Skrewness Reduction:



Log(1 + x)

Standardization

$$z = \frac{x - \mu}{\sigma}$$

μ = Mean

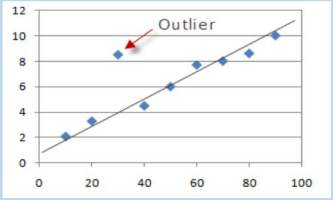
σ = Standard Deviation

Áp dụng phân phối chuẩn cho việc chuẩn hóa dữ liệu.

Features Engineering:

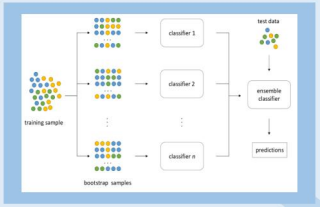
Loại dữ liệu	Phương thức
Tên tác giả, tên bài hát, Id	- Drop
Dữ liệu xiên (skew)	- Loudness áp dụng power transformer vì dữ liệu âm, - Còn lại áp dụng Log1P.
Dữ liệu phân phối chuẩn	Standartlize Scaler
Dữ liệu rời rạc	

Outlier Detection



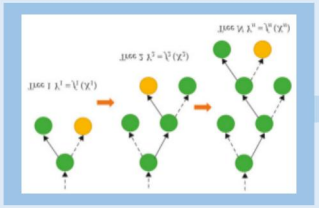
Model Training

Proposed Classification models:

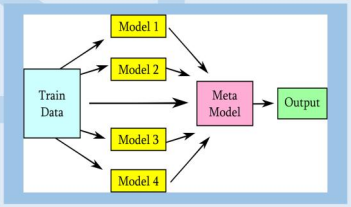


-XGB - Classifier

- Random Forest



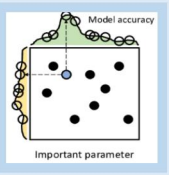
- Gradient boosting



Stacked Model

Tuning Method:

- Kết hợp giữa K-Fold và RandomSearch Cho các hyperparameters của các model.



Model Evaluating

- K-Fold Cross validation by Mean Square Error Evluating.

Models	Baseline		Optimized	Kaggle
Scoring	Accuracy	F1-Score	Accuracy	Accuracy
Gradient Boosting	0.5365	0.56	0.5452	0.53611
Random Forest	0.511	0.5599	0.5332	0.5333
XGB - Classifier	0.509	0.5655	0.5367	0.53611
Stacked Model (Optimized)				0.54259

Result