September 26, 2025

Mapping Musical Mood with Unsupervised Learning: PCA Spaces and Cosine-Similarity Recommendations

Kaleb Mercado[1], Claire Chang[2] (senior author)

[1] STISD World Scholars High School, Edinburg, Texas
[2] Cornell University, Ithaca, New York

Student Authors
Kaleb Mercado (high school)

SUMMARY

Music reliably evokes emotion, yet it is unclear how far we can model that response with lightweight, explainable machine learning. This study asks whether a system can recognize a song's mood and surface emotionally similar music without using raw audio pipelines or listener ratings. Using the Coimbra MIR group's 4Q dataset annotated by Russell's Circumplex Model of Affect, I merged musically meaningful features that summarize tempo, timbre, rhythm, and dynamics, along with tag encodings. I applied principal component analysis (PCA) to create a compact embedding; loadings suggested PC1 tracked dynamics and meter steadiness, and PC2 tracked rhythmic variability. The two-dimensional map aligned with four emotional quadrants: happy, angry, sad, relaxed. A cosine-similarity recommender retrieved nearest neighbors in this space and optionally emphasized songs near quadrant boundaries to reveal blended emotions. Unsupervised quality was characterized with scree and reconstruction-error curves and with clustering indices: silhouette, Davies-Bouldin, Calinski-Harabasz. Results showed coherent quadrant structure, clear elbows in dimensionality, and face-valid similarity groups. Because the approach remains in a tabular feature space, it is transparent and fast, with interpretable levers such as feature weights and tag contributions. This framework demonstrates a practical path toward mood-aware recommendations using explainable methods and publicly available features.

INTRODUCTION

Music's capacity to induce, modulate, and communicate emotion motivates research across psychology, neuroscience, and computer science. Modern streaming platforms increasingly organize catalogs by mood, which raises two linked questions: which musical attributes align with human affect, and can we model those relationships with methods that are simple to explain and easy to tune? Deep audio models can learn powerful representations, yet they require heavy pipelines and can be difficult to interpret. I pursue a complementary route that uses hand-crafted, musically interpretable features with classical unsupervised learning. Russell's circumplex model of affect positions emotions on two continuous axes: arousal (energy) and valence (positivity versus negativity). The two dimensional space supports practical categorization into quadrants: high arousal with positive valence (happy), high arousal with negative valence (angry or tense), low arousal with negative valence (sad), and low arousal with positive valence (relaxed) [3]. The Coimbra 4Q dataset supplies arousal and valence annotations with corresponding quadrant labels, which enables evaluation of unsupervised structure without training a classifier on the labels [1,2].

The pipeline has four steps. (1) Assemble a tabular matrix from dataset CSVs: numerical descriptors of tempo, timbre, rhythm, and dynamics, plus categorical tags for moods and genres. (2) Normalize features and apply PCA to uncover low dimensional structure. (3) Assess geometry with scree and reconstruction error curves and with clustering indices that use the known quadrants for evaluation. (4) Implement a cosine-similarity recommender in the learned space and examine retrieved neighbors for a seed song, including an option to emphasize boundary regions to surface blended emotions. I hypothesized that: (1) PCA on curated features would produce axes that align with interpretable musical dimensions, such as dynamics and rhythmic complexity, and that these axes would organize songs into regions corresponding to Russell's quadrants; and (2) cosine similarity in this space would retrieve emotionally coherent neighbors, including boundary cases that blend quadrant traits. The PCA map displayed quadrant coherence with strong clustering indices (silhouette 0.609, Davies–Bouldin 0.483, Calinski–Harabasz 3661.9). Scree and reconstruction-error curves justified a compact embedding. Cosine neighbors formed musically and emotionally plausible sets, including boundary cases. An explainable, tabular approach can serve as a credible backbone for mood-aware recommendations, complementing heavier audio-based systems.

RESULTS

Dataset and features. I used the Coimbra MIR 4Q dataset (900 clips) annotated with arousal, valence, and quadrant labels [1,2]. Four CSVs were merged into a single matrix by clip identifier. The final table retained 105 columns: about one hundred numeric audio descriptors and a compact set of tag encodings.

Tag encodings used in the model. To incorporate categorical information without inflating dimensionality, I used summary encodings rather than full one-hot vectors: **MoodsTotal** (count of all mood tags), **Moods** (count of mood tags matching the Warriner lexicon), **Genres** (count of genres), and **Sample** (binary indicator that a preview sample exists). **PQuad** was computed only for evaluation and was not included in the modeling matrix. The string lists (**MoodsFoundStr**, **MoodsStr**, **MoodsStrSplit**, **GenresStr**) and **SampleURL** served as metadata only and were not expanded in the model. All tag columns were z-scored with the audio features so that no single block dominated variance. These tags provide categorical context that listeners expect to influence recommendations.

Dimensionality structure. The scree plot showed a steep drop in explained variance across the first six to eight components, followed by a gradual tail (Figure 3). The cumulative curve exceeded 95% by roughly fifteen components, indicating diminishing returns beyond that range. Reconstruction mean squared error, computed after projecting to the first k components and reconstructing, flattened after about eight to ten components (Figure 2). After standardizing all feature blocks, no single component reached 90% explained variance; six to eight components were typically required, in line with the scree and reconstruction trends. Together, these diagnostics supported a compact embedding for visualization and retrieval.

Quadrant structure as clusters. Without using labels during PCA, the two-dimensional projection aligned with quadrant regions (Figure 1). Cluster quality metrics indicated good separation: silhouette 0.609, Davies–Bouldin 0.483, Calinski–Harabasz 3661.9. Lower Davies–Bouldin and higher silhouette and Calinski–Harabasz indicate more compact and well separated groupings, consistent with the visual quadrant layout.

Similarity recommendations. The recommender computes pairwise cosine similarity over normalized feature vectors and returns the top k neighbors after excluding the seed. Unless otherwise noted, k = 5. Candidates can optionally be filtered to points near quadrant boundaries by selecting songs within a narrow margin of a boundary in the PCA plane before ranking by cosine similarity. In the current figure set, the seed was “Dreams” by Fleetwood Mac; the retrieved neighbors share stylistic or affective traits consistent with their PCA locations (Table 1; Figure 1 for map context). This geometry also exposes boundary cases that blend traits from adjacent quadrants, which can support smooth mood transitions in playlists.

DISCUSSION

PCA1 loaded on dynamics and metrical steadiness; PCA2 loaded on rhythmic variability. These axes form an intuitive cross-section of musical organization that is consistent with a circumplex view of affect. Energetic, steady grooves tend to cluster in high-arousal regions, while softer dynamics and slower, regular patterns cluster in relaxed or sad regions. Supervised classification is possible because quadrant labels exist; The goal here is different: surface neighbors that feel emotionally coherent, reveal blends near boundaries, and allow user-controlled weights. An unsupervised geometry supports exploratory search and explainability. Users can inspect axis loadings and adjust feature contributions without retraining a classifier. Scree and reconstruction curves indicated that six to ten components captured structure. Apparent discrepancies across early exports were due to preprocessing choices: whether tag blocks were standardized before concatenation and the relative weight of high-variance categorical encodings. After z-scoring all feature blocks and limiting the influence of high-cardinality tags, elbows stabilized around 6-8 components and component interpretations were consistent. Emotion is subjective, so the system captures structural similarity rather than listener-specific nuance. Two common outliers were tracks with atypical production that confound loudness features and tracks whose tags conflict with audio descriptors. Future work will tune feature weights with small listening tests, add boundary-aware sampling that targets mixed-emotion regions, and explore semi-supervised objectives that nudge the geometry toward quadrant labels while preserving interpretability.

MATERIALS AND METHODS

Code availability. Code is at: https://github.com/Kkongmerc/Emotion-Based-Music-Model

Dataset. University of Coimbra MIR "4Q Audio Emotion" dataset annotated with arousal, valence, and quadrant labels, with companion CSVs for metadata, annotations, feature values, and feature descriptions [1,2].

Features and preprocessing. Four CSVs were merged by clip identifier. The final table included 105 columns: numeric descriptors of tempo, timbre, rhythm, loudness, and dynamics, plus tag encodings MoodsTotal, Moods, Genres, and Sample. PQuad was computed only for evaluation and was not included in the modeling matrix. Missing numeric values were imputed with column means; multi-hot vectors were filled with zeros. All numeric columns were z-scored prior to PCA and cosine similarity. Scalar multipliers were optionally applied to tag blocks to limit dominance by high-variance one-hot columns.

Dimensionality reduction. PCA was fit with scikit-learn's implementation using default parameters and a fixed random state [7]. Model selection used (i) the scree curve of explained-variance ratios and (ii) reconstruction mean-squared error after projecting to the first k components and reconstructing.

Cluster quality. PCA scores (PCA1 versus PCA2) were plotted with dashed quadrant boundaries derived from arousal and valence thresholds. Treating quadrant labels as clusters, I computed the silhouette coefficient, the Davies-Bouldin index, and the Calinski-Harabasz index using scikit learn.

Recommendation function. For a seed song, I computed cosine similarity between the seed vector and all other songs, then returned the top neighbors after excluding the seed. An optional boundary filter restricted candidates to a small margin around a quadrant boundary in the PCA plane.

Software. Python, pandas, scikit-learn, matplotlib, seaborn.

ACKNOWLEDGMENTS (Optional)

I thank the Coimbra MIR group for curating and sharing a richly annotated dataset.

REFERENCES

1. Panda, R., Malheiro, R., and Paiva, R. P. “Novel Audio Features for Music Emotion Recognition.” IEEE Transactions on Affective Computing (2018, early access). doi:10.1109/TAFFC.2018.2820691.

2. Panda, R., Malheiro, R., and Paiva, R. P. “Musical Texture and Expressivity Features for Music Emotion Recognition.” Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR 2018), Paris, France (2018).

3. Russell, J. A. “A Circumplex Model of Affect.” Journal of Personality and Social Psychology 39 (1980): 1161–1178.

4. Rousseeuw, P. J. “Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis.” Journal of Computational and Applied Mathematics 20 (1987): 53–65.

5. Davies, D. L., and Bouldin, D. W. “A Cluster Separation Measure.” IEEE Transactions on Pattern Analysis and Machine Intelligence 1.2 (1979): 224–227.

6. Caliński, T., and Harabasz, J. “A Dendrite Method for Cluster Analysis.” Communications in Statistics 3.1 (1974): 1–27.

7. Pedregosa, F., et al. “Scikit-learn: Machine Learning in Python.” Journal of Machine Learning Research 12 (2011): 2825–2830.

Figures and Figure Captions

Figure 1. PCA song map with Russell quadrants. Scatter of songs in PCA1 (Dynamics + Meter) vs PCA2 (Rhythmic Complexity) space. Dashed lines mark quadrant boundaries. Labeled points highlight example recommendations returned by cosine similarity for a chosen seed.

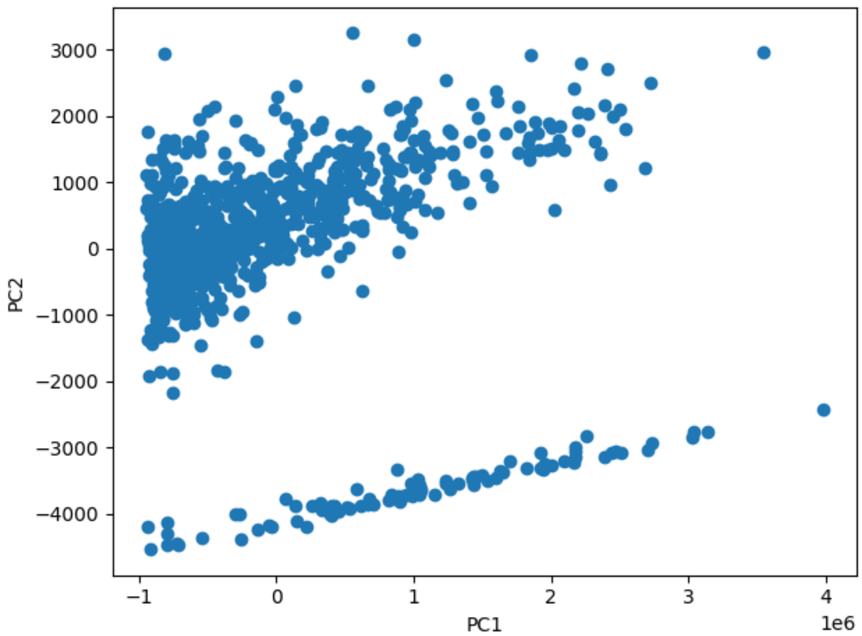

Figure 2. PCA reconstruction error vs. number of components. Mean-squared error from projecting to the first k components and reconstructing; curve flattens after ~8-10 components.

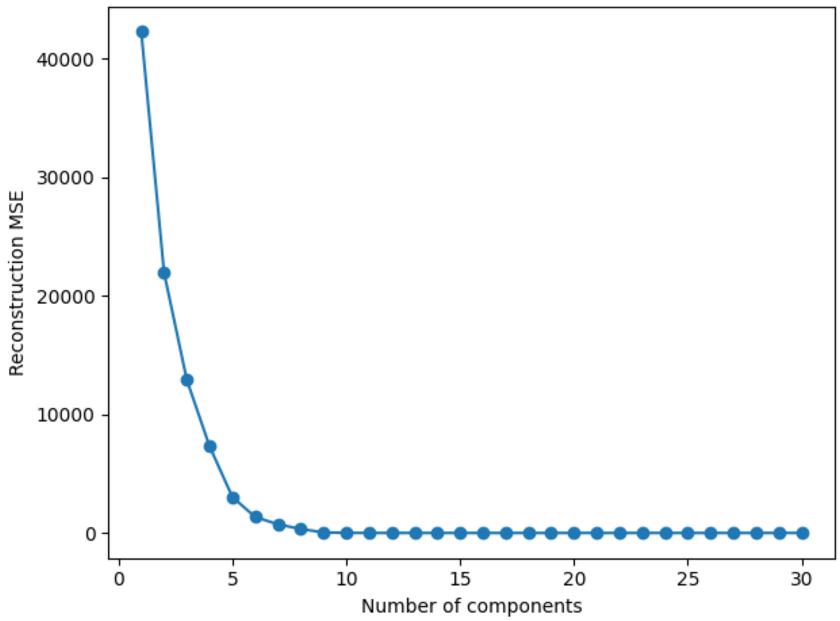

Figure 3. Scree plot (explained-variance ratio and cumulative variance). The elbow appears around 6-8 components; cumulative variance exceeds ~95% by ~15 components, with diminishing returns thereafter.

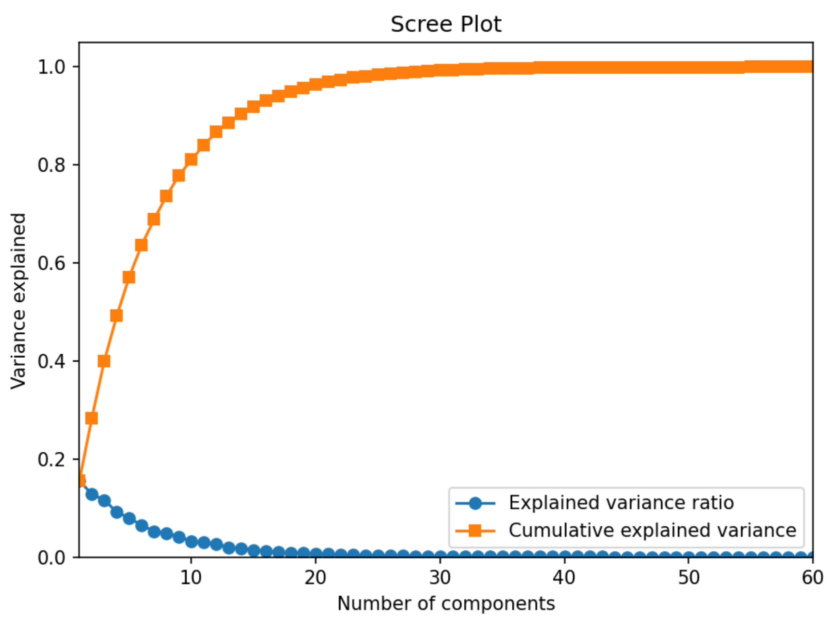

Figure 4. Emotionally blended recommendations for "Dreams" by Fleetwood Mac.

```
Selected Song: "Dreams" by Fleetwood Mac

Emotionally Blended Recommendations:
```

	Title	Artist	Quadrant_x	PCA1	PCA2	Similarity
0	Por Fin	Los Dandy's	Q4	-0.106	-3.001	0.779
1	St. Judy's Comet	Paul Simon	Q4	-5.202	0.107	0.730
2	Give It Up	Lil' Kim	Q4	-0.164	1.760	0.717
3	Beautiful Boy	Céline Dion	Q4	-2.750	0.009	0.649
4	What Kind of Fool (Do You Think I Am)	The Tams	Q4	1.077	-0.119	0.639

Table 1. Top five nearest neighbors returned by cosine similarity in the PCA feature space, showing each track's quadrant label, PCA1 and PCA2 coordinates, and cosine-similarity score. The seed "Dreams" by Fleetwood Mac is highlighted for context.