Bonus Exercise: Uncertainty in Machine Learning

Conformal Prediction with K-Nearest Neighbours

IMT Nord Europe — IC2 S2.1 AARES Project

Faycal Raghibi

February 2026

1 Introduction

Standard classification models produce a single predicted label for each input. While convenient, this point prediction carries no information about the model's confidence or the prediction's reliability. In many domains — medical diagnosis, autonomous driving, content moderation — understanding *how uncertain* a prediction is proves as important as the prediction itself.

Conformal prediction [1] is a framework that addresses this limitation. Rather than outputting a single label, it returns a *prediction set* — a subset of possible labels — that is guaranteed to contain the true label with a user-specified probability $1 - \varepsilon$. This guarantee is *distribution-free*: it holds without assumptions on the data distribution, requiring only that the training and test samples be exchangeable (e.g. i.i.d.).

In this report, we apply conformal prediction to the Spotify genre classification task using a K-Nearest Neighbours (KNN) classifier, following the procedure described in the project assignment.

2 Conformal Prediction: Principle

The core idea of conformal prediction is to measure how *conforming* a new example is with respect to the training data. Given a classification model and a candidate label y for a new input x, the method proceeds as follows:

1. **Augment** the training set with the tentative pair (x, y).
2. **Compute** a nonconformity score α_i for every sample i in the augmented set. This score quantifies how "unusual" the sample is relative to its neighbours. For KNN, the nonconformity measure is defined as:
$$\alpha_i \;=\; \frac{d_{\text{same}}(x_i)}{d_{\text{diff}}(x_i)},$$
where d_{same} is the distance to the nearest neighbour of the *same* class and d_{diff} is the distance to the nearest neighbour of a *different* class.
3. **Compute the p-value** for label y:
$$p(y) \;=\; \frac{\#\{i : \alpha_i \geq \alpha_{\text{new}}\}}{n+1}.$$
4. **Include** label y in the prediction set if and only if $p(y) \geq \varepsilon$.

The resulting prediction set satisfies:

$$\Pr\big[\, y_{\text{true}} \in \Gamma^{\varepsilon}(x) \,\big] \;\geq\; 1 - \varepsilon,$$

where $\Gamma^{\varepsilon}(x)$ denotes the prediction set at significance level ε.

3 Experimental Setup

3.1 Dataset

We use the Spotify training dataset (`spotify_dataset_train.csv`). As specified in the assignment, only the **first 1 500 samples** are retained in order to keep computation feasible — the full conformal procedure has $O(n \times C)$ complexity per test point, where n is the training size and C the number of classes.

3.2 Features

The following 14 numerical audio features were selected: `danceability`, `energy`, `key`, `loudness`, `mode`, `speechiness`, `acousticness`, `instrumentalness`, `liveness`, `valence`, `tempo`, `duration_ms`, `popularity`, and `explicit`. Missing values were imputed using the mean strategy, and all features were standardised to zero mean and unit variance.

3.3 Model

A **KNN classifier** ($k = 5$, Euclidean distance) from scikit-learn was trained on the 1 500-sample subset. The trained model was then passed to the `ConformalPrediction` class provided in `utils_projet.py`.

4 Experimental Protocol

For each test sample, the following steps were executed:

1. **Preprocessing:** The test sample was imputed and standardised using the same transformers fitted on the training subset.

2. **Conformal p-values:** The `predict(x_new)` method was called, computing one p-value per candidate genre.

3. **Interval at $\varepsilon = 0.05$:** The `compute_interval(0.05)` method returned all genres with $p \geq 0.05$, forming the 95%-confidence prediction set.

4. **Varying ε:** The interval was recomputed for $\varepsilon \in \{0.01, 0.05, 0.10, 0.20, 0.30, 0.50, 0.70, 0.90\}$ to observe how the set size evolves.

5. **Single-genre threshold:** ε was increased incrementally until the prediction set contained exactly one genre, yielding a point prediction with quantified confidence.

6. **Multi-sample evaluation:** Steps 1–5 were repeated for several test samples to assess variability.

5 Results

5.1 Interval Size vs. Significance Level

Table 1 summarises the typical relationship between ε and the prediction set size, observed across test samples.

ε	**Confidence**	**Avg. Set Size**
0.01	99%	$\approx$ 16–17
0.05	95%	$\approx$ 13–15
0.10	90%	$\approx$ 11–13
0.20	80%	$\approx$ 8–11
0.30	70%	$\approx$ 6–9
0.50	50%	$\approx$ 3–6
0.70	30%	$\approx$ 1–3
0.90	10%	$\approx$ 0–1

Table 1: Prediction set size as a function of the significance level ε. Exact values depend on the test sample.

The trend is clear: **the prediction set shrinks monotonically as ε increases**. At $\varepsilon = 0.05$ (95% confidence), the set typically contains 13–15 genres out of approximately 17 total — a wide interval that reflects the intrinsic difficulty of genre separation based on audio features alone.

5.2 Genre Similarity within Intervals

Examining the genres that remain in the prediction set at moderate ε (e.g. 0.30–0.50) reveals that they tend to be musically related. For instance, genres such as *rock*, *alternative*, and *indie* often co-occur, as do *hip-hop* and *R&B*. This indicates that the nonconformity measure captures meaningful feature-space distances: genres with similar audio profiles produce similar p-values, keeping them in the prediction set together.

5.3 Single-Genre Predictions

For most test samples, the prediction set reduces to a single genre at ε values between 0.50 and 0.80, corresponding to confidence levels of 20–50%. In some cases, the single-genre threshold is even higher, reflecting particularly ambiguous samples.

The genre identified at the single-genre threshold generally agrees with the KNN point prediction but comes with an explicit confidence estimate — a significant advantage over a bare point prediction.

6 Discussion and Conclusions

6.1 Interpretation

The conformal prediction framework transforms a KNN classifier into a calibrated uncertainty-aware predictor. The key observations are:

- **Coverage guarantee:** The prediction set at significance ε is guaranteed to contain the true label with probability $\geq 1 - \varepsilon$. This holds without any distributional assumptions beyond exchangeability.
- **Wide intervals reflect model limitations:** The large prediction sets at $\varepsilon = 0.05$ are consistent with the modest F1 score (≈ 0.44) achieved by classic classifiers on this dataset. Audio features alone cannot reliably distinguish many genre pairs.
- **Practical trade-off:** A practitioner must choose between a broad, high-confidence set (small ε) and a narrow, lower-confidence set (large ε). The appropriate choice depends on the downstream application.

6.2 Limitations

- **Computational cost:** Full conformal prediction is expensive ($O(n \times C)$ per test point), limiting scalability. Inductive (split) conformal prediction could reduce this cost by separating the calibration step.
- **Training size:** Restricting to 1 500 samples reduces the KNN's accuracy, which in turn inflates the prediction sets.
- **Model choice:** KNN is sensitive to the curse of dimensionality. Alternative classifiers (e.g. Random Forests, gradient-boosted trees) with adapted nonconformity measures might produce tighter intervals.

6.3 Conclusion

Conformal prediction provides a rigorous, distribution-free method for quantifying prediction uncertainty. Applied to genre classification with KNN, it reveals the inherent difficulty of the task: reliable predictions require broad genre sets, and narrow (single-genre) predictions come with correspondingly reduced confidence. This exercise highlights the importance of uncertainty quantification as a complement to point predictions in practical machine learning systems.

References

[1] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.