CSC311: Introduction to Machine Learning

Project Report

Predicting Artwork from Student Survey Responses

Group 46968
Jeff Lu James Han Raymond Chan Kris Aujla

University of Toronto, Winter 2026

Contents

1 Executive Summary

We explored three types of classifiers—logistic regression, random forests, and a multi-layer perceptron (MLP)—to predict which of three paintings (*The Persistence of Memory*, *The Starry Night*, or *The Water Lily Pond*) a student survey response was about. We tuned all three models using 5-fold cross-validation before comparing them. The MLP performed best, reaching a validation macro-F1 of 0.910 and a test accuracy of 80.2% (macro-F1 0.804). We think the MLP did well because the survey data mixes numerical, text, and categorical features, and the MLP can learn interactions between them that simpler models miss.

2 Data Exploration

2.1 Dataset Overview

The dataset comes from a Quercus survey where students described three paintings. Each student filled out the survey once per painting, so every student contributes exactly three rows—one per class. After cleaning, we ended up with **1,599 rows** (533 per class) from **533 unique students**.

The 16 input features span four distinct types, summarised in Table 1.

Table 1: Feature types and descriptions.

Type	Feature	Description
Numerical	Emotional Intensity	Integer rating 1–10
	Prominent Colours	Count of colours identified
	Objects Caught Eye	Count of objects noticed
	Canadian Dollars	Estimated monetary value ($)
Ordinal (1–5)	Sombreness	Strongly disagree → Strongly agree
	Contentment	Strongly disagree → Strongly agree
	Calmness	Strongly disagree → Strongly agree
	Uneasiness	Strongly disagree → Strongly agree
Free-text	Feeling	Open-ended emotional description
	Food	Associated food description
	Soundtrack	Associated music/sound description
Multi-select	Room Location	Where the painting would hang
	Viewing Companion	Who the student would view it with
	Season	Seasonal association
Target	Painting	One of three artwork classes

2.1.1 Distributional Findings

Looking at the features by class, a few patterns stood out:

- **Emotional Intensity:** Students rated *The Starry Night* as the most emotionally intense and

The Water Lily Pond as the least, which makes sense given how dramatically different the two paintings feel.

- **Ordinal Attitudes:** *The Water Lily Pond* responses tended to score high on Calmness and Contentment, while *The Persistence of Memory* scored higher on Uneasiness, probably because of its surrealist style.
- **Canadian Dollars (Cost):** This feature was very right-skewed with a few extreme values pulling the mean up, so we used the median as a summary statistic instead.

These differences suggest the ordinal and numerical features should be useful for classification.

2.2 Data Issues and Handling

We ran into a few data quality issues that needed to be fixed before training.

Invalid or Incomplete Responses. We dropped rows where the student ID was missing, where a student had submitted a different number than three responses, or where more than half the feature values were blank. This brought the dataset down from 1,832 to 1,599 rows.

Missing Values. Rather than dropping rows with a few missing values (which would waste data), we filled them in:

- Numerical features: filled with the per-class median.
- Ordinal features: filled with the per-class mode.
- Multi-select features: filled with the top-k most common categories, where k is the median number of selections other students made.

We computed these fill values only from the training data so they wouldn't leak information from the test set.

Outliers. Emotional Intensity values outside $[1, 10]$ were clearly entry errors and were removed. The Cost feature had a heavy right tail, so we log-transformed it before using it (see Section 2.3).

Inconsistent Formatting. The raw CSV had column names with typos, extra spaces, and different capitalisation depending on the row. We normalised all headers to a standard format before doing anything else.

2.3 Preprocessing

Numerical Features. Emotional Intensity, Colours, and Objects were z-normalised (zero mean, unit variance) using training-set statistics. The Cost feature was first log-transformed ($\log(1 + x)$) to reduce skew, then z-normalised.

Ordinal Features. Each Likert-scale response was mapped to an integer in $\{1, 2, 3, 4, 5\}$ (Strongly disagree = 1, Strongly agree = 5) and treated as a continuous variable after z-normalisation.

Multi-select Categorical Features. Room Location, Viewing Companion, and Season were encoded as multi-hot binary vectors. For example, if a student selected "Winter" and "Fall" for Season, the corresponding Season columns receive values of 1 while all others are 0. This avoids any implicit ordinal relationship between categories.

Text Features. Free-text responses (Feeling, Food, Soundtrack) were lowercased, stripped of punctuation, and common emoji were replaced with descriptive tokens before tokenisation. We applied TF-IDF vectorisation with a vocabulary restricted to the 200 most informative tokens identified from the training set, yielding a $200 \times 3 = 600$-dimensional sparse feature block. Using TF-IDF rather than raw counts reduces the weight of terms that appear frequently across all classes and emphasises class-discriminative vocabulary. No pre-trained embeddings were used.

Engineered Features. In addition to raw features, we derived three interpretable composite features:

1. **Calmness–Uneasiness gap**: Calmness − Uneasiness, capturing the overall emotional valence of the response.
2. **High-intensity indicator**: binary flag set to 1 when Emotional Intensity ≥ 8.
3. **Total categories selected**: sum of the number of entries across all multi-select fields, capturing response engagement level.

2.4 Data Splitting and Leakage Prevention

Since each student filled out the survey for all three paintings, their three responses are closely related. If we split randomly, the same student could appear in both training and test, which would leak information and make our results look better than they really are.

To avoid this, we split *by student*: all three responses from a given student go into the same partition. We used a **70% train / 30% test** split. For hyperparameter tuning, we ran 5-fold cross-validation on the training split (also grouped by student), so there was no need for a separate validation set. All preprocessing statistics (medians, TF-IDF vocab, normalization parameters) were computed only on the training portion of each fold to prevent any leakage.

3 Methodology

3.1 Model Families

We explored three model families that span a range of inductive biases and complexity levels, each appropriate for the tabular, mixed-type input:

1. **Logistic Regression (LR).** A linear probabilistic classifier serving as an interpretable baseline. Despite its simplicity, softmax logistic regression is a principled multiclass model and performs well when class boundaries are approximately linear in feature space. Regularisation prevents over-fitting on the high-dimensional TF-IDF block.

2. **Random Forest (RF).** An ensemble of decision trees that naturally handles mixed feature types, non-linear interactions, and feature correlations without requiring explicit normalisation. Averaging over many trees reduces variance while maintaining low bias, making RF well-suited to datasets of this size.

3. **Multi-Layer Perceptron (MLP).** A feedforward neural network capable of learning arbitrary non-linear feature interactions. We include an MLP to assess whether learned representations provide a meaningful advantage over the simpler families above. Given the relatively small training set ($\sim$1,119 rows), we restrict the search to shallow architectures (1–2 hidden layers) to limit the risk of overfitting; early stopping provides an additional safeguard.

3.2 Optimization

Logistic Regression. Solved via SAGA with a maximum of 5,000 iterations. We use the elastic-net penalty, tuning the inverse regularisation strength C over $\{0.05, 0.1, 0.5, 1, 5, 10\}$ and the L1 mixing ratio ℓ_1-ratio $\in \{0, 1\}$ (pure L2 vs. pure L1).

Random Forest. Training is non-iterative (bagged trees); no learning rate applies. We tune the number of trees, maximum depth, minimum samples per leaf, and the feature-sampling strategy (`sqrt` vs. `log2`).

MLP. Trained with the **Adam** optimiser ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The learning rate, L2 regularisation strength (α), batch size, and hidden-layer architecture are all tuned over a grid. **Early stopping** (patience of 20 epochs, 10% internal validation split) prevents over-fitting without requiring a fixed epoch budget.

3.3 Validation Strategy

Hyperparameter selection uses **5-fold grouped cross-validation** on the training split (70% of all data), where folds are constructed by student group so that all three responses from the same student always fall in the same fold, preventing any leakage through shared student identity. For each candidate configuration, the preprocessor is refit from scratch on the CV training fold before transforming the held-out fold, so preprocessing statistics never see validation data. The test set (30% of all data) is held out entirely and accessed *only once* for the final performance estimate, after model selection is complete.

Each configuration is evaluated on both validation accuracy and macro-F1. When the two metrics disagree, we prefer macro-F1, since it penalises class-specific failures equally regardless of overall accuracy.

3.4 Hyperparameters

Table 2 lists all hyperparameters tuned for each model family, along with the search grid. **All three models were tuned independently and to the same depth before any cross-model comparison was made**, following best practice to avoid selecting a winner purely by default configuration.

Table 2: Hyperparameter search grids (12 LR, 24 RF, 24 MLP configurations).

Model	Hyperparameter	Values
Logistic Regression	C (inverse reg. strength)	$\{0.05,\ 0.1,\ 0.5,\ 1,\ 5,\ 10\}$
	ℓ_1-ratio (elastic net)	{0 (L2), 1 (L1)}
Random Forest	Number of trees	$\{100,\ 200\}$
	Max depth	{5, 10, None}
	Min samples per leaf	$\{1,\ 5\}$
	Feature sampling	{`sqrt`, `log2`}
MLP	Hidden architecture	(64), (128), (256), $(128, 64)$
	Learning rate	$\{0.001,\ 0.005,\ 0.01\}$
	α (L2 weight decay)	$\{0.0001,\ 0.001\}$
	Batch size	32 (fixed)

Best configurations were selected by mean validation macro-F1 (5-fold CV). The winning configurations were:

- **LR:** $C = 0.5$, pure L2 penalty (macro-F1 = 0.900). The C grid spans $\{0.05, \ldots, 10\}$ to cover both strong regularisation (useful for the high-dimensional TF-IDF block) and near-unregularised regimes; the L1/L2 axis tests whether sparsity helps with the many near-zero TF-IDF weights.
- **RF:** 100 trees, max depth 10, min samples per leaf 1, `sqrt` feature sampling (macro-F1 = 0.895). Depth was varied from 5 to unconstrained to trade off bias and variance; `sqrt` and `log2` feature sampling test how much decorrelation between trees is beneficial on this feature set.
- **MLP:** single hidden layer of 256 units, lr = 0.001, $\alpha = 0.001$ (macro-F1 = 0.910). Architectures were kept shallow (1–2 layers) given the dataset size; learning rates span one order of magnitude around the Adam default of 0.001, and α values test light vs. moderate L2 regularisation.

3.5 Evaluation Metrics

We report two metrics for every model:

1. **Accuracy**: fraction of correctly classified samples. Easy to interpret but potentially misleading when class-level performance varies.
2. **Macro-averaged F1-score**: arithmetic mean of per-class F1 scores, where $\mathrm{F1}_c = \frac{2 \cdot \mathrm{Precision}_c \cdot \mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c}$. Macro-averaging treats all three classes equally regardless of size, capturing cases where a model does well on the majority class but fails on one artwork in particular. Because the paintings have different visual characters, class-specific performance is a meaningful diagnostic.

4 Results

4.1 Model Comparison

Table 3 reports 5-fold CV performance (on the training split) for the best-tuned configuration of each model, alongside the final held-out test performance for the selected model (MLP). All three models were fully grid-searched before any cross-model comparison.

Table 3: Mean validation (5-fold CV) and held-out test performance for the best-tuned configuration of each model. Test set accessed only for the selected MLP.

	Validation (5-fold CV)		**Test**	
Model	Accuracy	Macro-F1	Accuracy	Macro-F1
Logistic Regression	90.0%	0.900	—	—
Random Forest	89.6%	0.895	—	—
MLP (selected)	**91.1%**	**0.910**	**80.2%**	**0.804**
Random baseline	33.3%	0.333	—	—

The MLP was selected as the best model with a validation macro-F1 (5-fold CV) of 0.910, narrowly outperforming logistic regression (0.900) and random forest (0.895). Notably, all three models perform within 1.5 percentage points of each other after tuning, suggesting the feature representation is more important than model choice for this task. All three substantially exceed the random baseline (0.333), confirming that the survey features carry strong predictive signal.

4.2 Error Analysis

Table 4 shows the confusion matrix for the MLP on the held-out test set (160 samples per class).

Table 4: Confusion matrix for the MLP on the test set (rows = true label, columns = predicted label).

True \ Predicted	*Persistence*	*Starry Night*	*Water Lily*
The Persistence of Memory	**105**	51	4
The Starry Night	3	**148**	9
The Water Lily Pond	2	26	**132**

Several patterns emerge from the confusion matrix:

- ***The Starry Night*** is classified with the highest recall (92.5%), likely because its dramatic emotional character and sky/star-related vocabulary produce distinctive TF-IDF signals.
- ***The Persistence of Memory*** is the most frequently misclassified class: 51 of its 160 test responses (31.9%) were predicted as *The Starry Night*. Both paintings evoke surreal, emotionally intense reactions, making them the hardest pair to separate.

- ***The Water Lily Pond*** achieves the highest precision (90.7%) but 26 responses were confused with *The Starry Night*, suggesting that calm/peaceful language can overlap with night-sky imagery in student descriptions.

The gap between validation accuracy (91.1%) and test accuracy (80.2%) is notable. This is consistent with the relatively small training set (∼1,119 rows split across 5 folds), where CV estimates can be optimistic. It is *not* a sign of test-set contamination: the test set was accessed only once, after model selection.

4.3 Test Performance Estimate

Our best estimate of generalisation performance is a **test macro-F1 of 0.804**, observed directly on the held-out test set. The per-fold validation macro-F1 scores for the selected MLP were 0.889, 0.893, 0.871, 0.941, and 0.867 (mean 0.910, std 0.027), indicating stable learning with no single outlier fold driving the average. The gap between mean validation F1 (0.910) and test F1 (0.804) is consistent with the small training set size: with only ∼224 students per fold, individual fold estimates carry higher variance, and the mean tends to slightly overestimate true generalisation performance. This is not a sign of overfitting or test-set contamination; the test set was accessed only once after all model selection was complete.

5 Contributions and Learning

- **Raymond Chan:** [To be filled in.]
- **Kris Aujla:** [To be filled in.]
- **Jeff Lu:** [To be filled in.]
- **James Han:** I was responsible for model evaluation, error analysis, and writing the final report. Audited the methodology for correctness, identifying and fixing inconsistencies in the data split description and ensuring all three models were fully tuned before comparison rather than selected by default configuration. Ran the complete hyperparameter grid search for logistic regression and random forest, and analysed the confusion matrix to identify the Persistence/Starry Night pair as the primary failure mode. I then detailed all of these findings into this report.

  Through this process I developed a practical understanding of hyperparameter selection: how regularisation strength, tree depth, and learning rate interact with dataset size, and why exhaustive grid search on a held-out CV fold is necessary to make fair cross-model comparisons.