A logo for college computing

Description automatically generated

**Assessment Cover Page**

|  |  |
| --- | --- |
| *Student Full Name* | Isabella Gubitoso |
| *Student Number* | sba25155 |
| *Module Title* | Machine learning |
| *Assessment Title* | CA1 – Machine Learning |
| *Lecturer/Supervisor* | Dr. Muhammad lgbal |
| *Assessment Due Date* | 25/11/2025 |
| *Date of Submission* | 30/11/2025 |

**Use of AI Tools**

I have not used any AI tools or technologies in the preparation of this assessment.

**Declaration**By submitting this assessment, I confirm that I have read the CCT policy on academic misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source.

I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Contents

[Introduction 1](#_heading=h.gjdgxs)

[1](#_heading=h.30j0zll) Data Characterization and Preprocessing 2

2 Hyperparameter Tuning 3

3 Results 4

4 Conclusion5

[References 6](#_heading=h.2et92p0)

# Introduction

This assignment explores price classification within the Arts, Culture and Heritage domain using an art e-commerce dataset. The motivation for this topic stems from the expanding digital art marketplace and the inherent challenges sellers face when pricing their artworks in a market characterized by significant subjective valuation

The dataset comprises information on 2,500 paintings sold through an online platform, encompassing features such as painter identification, artistic style, medium employed, dimensions, framing specifications, and various aesthetic characteristics. The primary objective is to predict price categorization based on these available features.

This problem presents particular interest because, unlike conventional e-commerce products, art pricing incorporates both subjective elements and objective factors. Understanding which features exert the strongest influence on price categories could provide valuable insights for sellers in pricing strategies and help buyers comprehend value drivers in online art marketplaces.

The research approach treats this as a classification problem employing two distinct machine learning models: Random Forest and Support Vector Machines (SVM). These models were selected due to their demonstrated effectiveness in handling categorical data and their capacity to capture complex, non-linear relationships between features (Muller and Guido, 2016).

Word count: 186

# 1 Data Characterization and Preprocessing

The original dataset contained 2,500 records across 18 columns. Initial data exploration revealed that the 'Reproduction Type' column exhibited 677 missing values, representing approximately 27% of the data. Rather than eliminating these rows and reducing the dataset size substantially, missing values were imputed with 'Not Applicable', a reasonable categorical assignment for artworks where reproduction type is not relevant.

Also, the 'Delivery (days)' column was removed from analysis as delivery timeframes represent logistical characteristics rather than product attributes that would influence inherent price categorization.

For target variable creation, quantile-based discretization was employed to transform continuous price values into three categories. This methodology ensured class balance: Low (34%), Medium (32.68%), and High (33.32%). Maintaining balanced classes is crucial as it prevents model bias toward overrepresented categories during training.

As per assignment requirements, three distinct train-test splits were evaluated: 10%, 15%, and 20% test set proportions. For each configuration, both Random Forest and SVM models were trained and evaluated using 5-fold cross-validation on the training data to ensure robust performance estimation.

Results across the three splits demonstrated notable patterns. Random Forest builds multiple decision trees and combines their predictions, which reduces overfitting and improves accuracy (Müller and Guido, 2016). For this assessment, Random Forest achieved optimal performance with the 10% test split, attaining 41% test accuracy with a cross-validation mean of 32.53%. As test set size increased, test accuracy decreased moderately (38% for 15% split, 35% for 20% split), suggesting that larger training sets provide performance benefits, though cross-validation scores remained relatively stable around 31-33%.

Burkov (2019) explains the Support Vector Machine seeks to optimal hyperplane that maximizes the margin between classes, contributing to better generalization capability of the model. For this project, SVM exhibited more consistent performance across splits, with test accuracies ranging from 32.8% to 35.2% and cross-validation means approximating 30-31%. Cross-validation standard deviations for both models remained relatively small, indicating reasonably stable performance across folds.

Word count: 315

# 2 Hyperparameter Tuning

Hyperparameter tuning constitutes an essential component of machine learning model optimization, as it facilitates identification of optimal parameter configurations that are not learned during the standard training process (Burkov, 2019). This systematic approach ensures models operate at their maximum potential given the available data.

For this analysis, GridSearchCV with 3-fold cross-validation was implemented on the 20% test split configuration to systematically evaluate different hyperparameter combinations for both models. Burkov (2019) emphasizes that the training process of machine learning models essentially consists of solving optimization problems, seeking paramenters that minimize prediction error.

Following GridSearchCV optimization with 3-fold cross-validation, both models identified optimal parameter configurations:

Random Forest optimal parameters:

- n\_estimators: 300

- max\_depth: 20

- min\_samples\_split: 2

- Cross-validation score: 32.90%

SVM optimal parameters:

- C: 0.1

- gamma: scale

- kernel: rbf

- Cross-validation score: 34.50%

The Random Forest configuration achieved test accuracy of 35.6%, while SVM achieved 32.0%.

The SVM model similarly underwent systematic hyperparameter search, identifying optimal parameters though maintaining performance around 32-35% accuracy. A comprehensive comparison table was generated displaying accuracy, precision, recall, and F1 scores for both optimized models, facilitating direct performance comparison. Random Forest was selected as the first classification algorithm because it works well with categorical data and handles multiple features effectively. Random Forest builds multiple decision trees and combines their predictions, which reduces overfitting and improves accuracy. It is also useful for this problem because it can identify which features are most important for predicting price categories.

Table 1



The Random Forest model was configured with 100 trees and a fixed random state for reproducibility. It was trained and tested across three different data splits to evaluate consistency.

Word count: 273

3 Results

The obtained results reveal that both models achieved accuracy levels of only 30-40%, which initially appears suboptimal. However, contextual analysis is necessary: with three balanced classes, random classification would yield approximately 33% accuracy. The models therefore demonstrate only marginal improvement over chance-level performance.

Examination of confusion matrices indicates that both models struggled to discriminate effectively among the three price categories. This suggests that the available features may not constitute strong predictors of price. Art pricing likely depends heavily on factors absent from this dataset, including artist reputation, artwork provenance, historical significance, and current market trends.

Regarding overfitting and underfitting considerations, the models do not exhibit severe overfitting, as cross-validation scores approximate test accuracies reasonably well. Substantial overfitting would manifest as significantly elevated training/cross-validation scores compared to test performance. However, the overall modest performance suggests possible underfitting.

Comparative analysis between Random Forest and SVM reveals Random Forest achieving marginally superior performance, which aligns with theoretical expectations given Random Forest's capability to model complex feature interactions through ensemble learning. Nevertheless, neither model achieved performance levels that would support practical deployment.

One positive finding is the relative stability of model performance across different train-test splits, suggesting that results are consistent and reproducible despite modest accuracy levels.

Word count: 203

# 4 Conclusion

This assignment provided comprehensive exposure to the complete machine learning pipeline, from initial data exploration through final model evaluation. While achieved accuracy levels fell below initial expectations, the process yielded valuable insights regarding the challenges inherent in predicting art prices based solely on aesthetic and categorical features.

This work demonstrates that machine learning, while powerful, requires adequate signal within the data. When underlying features lack sufficient information content, even sophisticated algorithms cannot generate accurate predictions, an important consideration when approaching real-world problems across any domain.

Word count: 85

# References

Burkov, A. (2019) The hundred-page machine learning book. Quebec City: Andriy Burkov.

Kaggle (2024) tArt Market Dataset: Selling Painting Predictions. Available at: https://www.kaggle.com/datasets/[SEU-URL-EXATO]https://www.kaggle.com/datasets/jijagallery/art-market-dataset-selling-paintings-prediction/data (Accessed: 15 November 2025).

Muller, A.C. and Guido, S. (2016) Introduction to machine learning with Python: a guide for data scientists. Sebastopol: O'Reilly Media.