Abstract

Background: Paintings differ considerably based on the style. There are four painting styles that are historically the most recognizable: impressionism, realism, surrealism, and landscape. A person's characteristics, lifestyle, or demographics may or may not have an influence on what style of painting they prefer. We analyzed a survey which paired paintings of different styles against each other. Four different painting style matchups were pitted against each other: "RA" – Landscape vs Surrealism, "MF" – Impressionisms vs Landscape, "LP" – Impressionism vs Realism, and "V" – Realism vs Landscape. During the survey, two random paintings appeared on the screen along with a 5-point Likert scale which measured the subjects' preference toward either painting. Along with the Likert scale responses, demographic and lifestyle characteristics of participants were also collected.
Methods: Invalid, duplicate, and incomplete survey responses were omitted from the dataset. The Likert-scale data was preprocessed in order to complete an EDA. Following the EDA, K Nearest Neighbor (KNN) and Random Forest (RF) classifications were done on the Likert scale responses (features) and the demographics (target).
Results: Both the KNN and RF classification outputs identified marriage status (specifically participants that were 'never married') as the best predictor (KNN accuracy= ~80% for all matchups after K=5; RF accuracy= ~80%). The KNN model performed terribly (0% correct) in predicting participants who were 'Previously married'. The model doesn't predict a value of 'currently married' correctly either. This can probably be attributed to the distribution of responses which are skewed heavily to 'never married'. The RF model showed extremely similar results as the KNN model, although it may be worth nothing that the accuracy of predicting 'previously married' responses minimally increased from 0% to 1%.
Conclusion: Due to the heavily skewed responses, this dataset is not recommended for any future KNN or RF classification analyses. It is my recommendation that future classification analyses would need to be done on data that is more evenly distributed, or on datasets with binary or dichotomous outcomes, not 4+ Likert scale outcomes.

1. Introduction

Some may argue that art is an incredibly important part of society. It is a way to express oneself and to connect, or disagree, with others based on similar or different interpretations. One of the oldest expressions of art in human history is painting. Paintings can differ drastically based on their subject(s) and style. When it comes to types of painting styles specifically, four are historically the most recognizable: impressionism, realism, surrealism, and landscape.

Impressionism is a 19th-century art movement characterized by relatively small, thin, yet visible brush strokes, open composition, emphasis on accurate depiction of light in its changing qualities (often accentuating the effects of the passage of time), ordinary subject matter, unusual visual angles, and inclusion of movement as a crucial element of human perception and experience. Impressionism originated with a group of Paris-based artists whose independent exhibitions brought them to prominence during the 1870s and 1880s [1]. Some famous impressionists are Camille Pissarro and Claude Monet.

Realism is an art style that focuses on making pieces look as realistic and true-to-life as possible. While the subjects may sometimes appear somewhat stylized, realism seeks to present subjects as they look in real life [2]. Rembrandt is considered the most famous Realist.

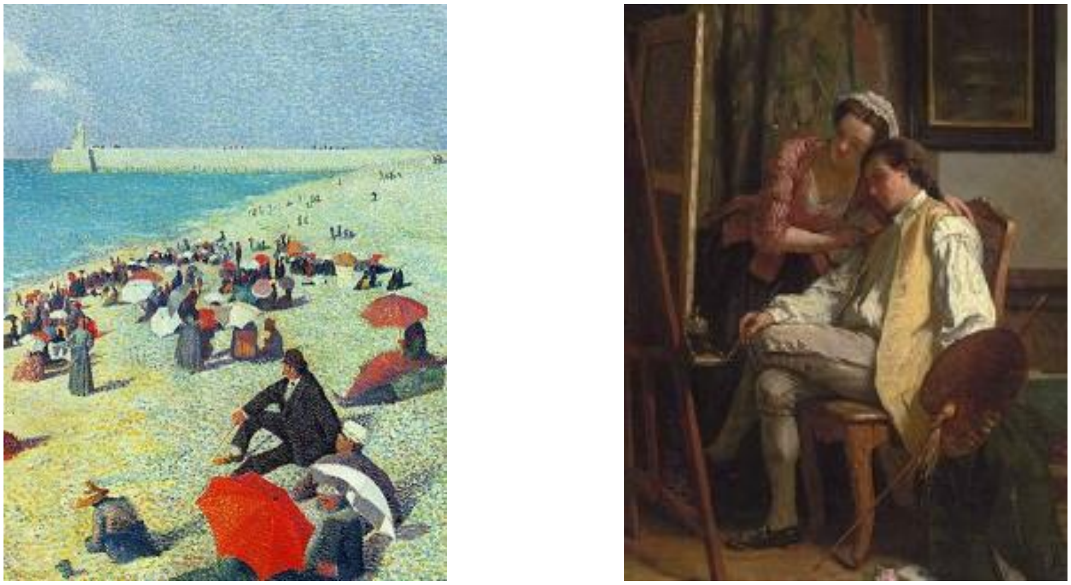
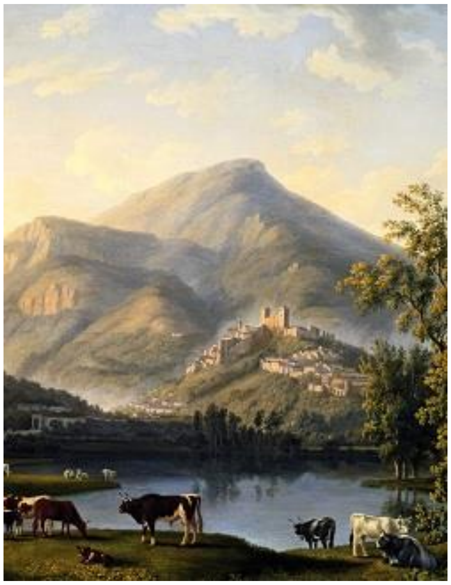


*Figure 1. Left, “*A Sunday Afternoon on the Island of La Grande Jatte” *by Léon Seurat, an example of Impressionism. Right, “*The Finishing Touches” by Jean Carlus, an Example of Realism, both in the “LP” matchup.

Much as the name of the style would suggest, Landscape painting is the depiction of natural scenery such as mountains, valleys, trees, rivers, and forests, especially where the main subject is a wide view—with its elements arranged into a coherent composition. Sky is almost always included in the view, and weather is often an element of the composition [3].

Finally, Surrealism is more of a movement than a style. It aims to revolutionize human experience. It balances a rational vision of life with one that asserts the power of the unconscious and dreams. The movement’s artists find magic and strange beauty in the unexpected and the uncanny, the disregarded and the unconventional [4]. The most famous surrealist is Salvador Dali.



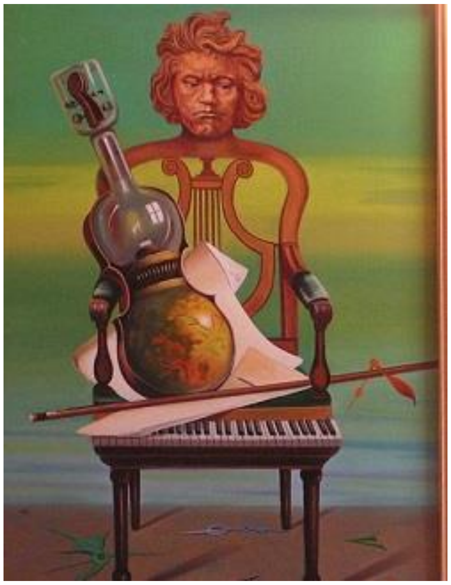


*Figure 2. Left, “Landscape with a View of Itri”, by Jakob Philipp Hackert, an example of Landscape style. Right, “Space and Time”, by William Girometti, an example of Surrealism. B*oth used in the “RA” matchup.

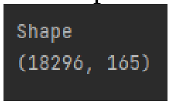
It’s fair to say that every person in the world has a unique preference in anything that is subjective, like music, cars, pets, furniture, etc. Having these unique preferences is what makes everyone an individual. Based on this, what if we were able to predict what style of painting a person preferred based on their characteristics, lifestyle, or demographics? Would we find that those with higher education prefer Abstraction? Do people who identify as asexual prefer Geometric art? Do those who practice Islam prefer Portraiture?

To test this out, a survey was created pairing paintings of different styles against each other. There were 4 different categories of paintings used: Surrealism, Impressionism, Realism, and Landscape. As well as 4 different matchups: “RA” – Landscape vs Surrealism, “MF” – Impressionisms vs Landscape, “LP” – Impressionism vs Realism, and “V” – Realism vs Landscape. When a user begins the survey, two paintings appear (randomly on right or left) on the screen from a matchup with 5 options below reading “Strongly prefer left” , “prefer left”, “prefer neither”, “prefer right”, and “Strongly prefer right”. This continues for all 31 paintings. After the user rates the paintings, there are several questions asked in order to find out gender, age, education level, if they grew up in an urban/suburban/rural environment, if they are right or left-handed, what religion they are, race, sexual orientation and so on. These things may subtly influence a person's preferences on their everyday life, unbeknownst to them.

2. Methods

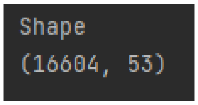
EDA & Preprocessing

Before we can really dive in and begin to look for trends in out data, we need to understand our data. Take a look to find trends at a high level before going into the nitty gritty. Taking a look at the shape of the data, we see that there are approximately 18,300 entries, and 165 columns.



```
Shape
(18296, 165)
```

That is a tremendous amount of entries, surely there are some trends in this data. However, 165 columns are surely too many and reduction is needed. For each matchup, a certain amount of data is collected. This data includes time elapsed to decide the user's preference, whether the picture is on the left or right, and the order of the matchups (which are random). While this might be useful information, there is simply an abundance of data here and we need to get rid of it. We will also omit data regarding the screen size of the user's computer, and the time it took to do all of the matchups and take the survey at the end. After these columns are removed, we are now down to 53 columns which is much more manageable as 30 of those are matchups. Next, we attempt to reduce the amount of columns if possible. The raw .csv file shows data that is incorrect, such as age = 2 and familysize (amount of immediate siblings) = 50. Outliers were removed from these columns only, as they were the only columns where the user could write in their answers. The remaining answers were chosen from a drop-down. Finally, entries that had a value of 'n/a' for any column were removed along with any duplicate entries. The data was reduced from ~18,300 entries to ~16,600 entries and 53 rows, which is much more manageable.



```
Shape
(16604, 53)
```

K Nearest Neighbor (KNN)

KNN is a type of supervised learning algorithm which is used for both regression and classification purposes, but mostly is used for classification problems [5]. KNN will try and predict the class of test data by calculating the distance between the test data and all of the training points. The 'K' is the amount of training points that are nearest to the data which will be used to classify the test point. Choosing the value of K is important as a lower K may lead to overfitting, while a K that is too high may lead to underfitting.

Random Forests (RF)

Random forests is a supervised learning algorithm. A forest is comprised of trees. It is said that the more trees it has, the more robust a forest is. Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting. It also provides a pretty good indicator of the feature importance [6].

3. Results

KNN

Below are four graphs, which show the accuracy of KNN with various targets and ‘K’ values. One thing to note is that the graphs become smoother with more K values. Additionally, it can be seen that these graphs look near identical. The accuracy for target ‘married’, is approximately 80% for all matchups (after K=5). Bringing up the rear in every case are ‘religion’, ‘education’, and ‘urban’. It’s clear to see that ‘married’ is performing the best in every case by a decent margin (~7% in every case). This will be our focus target going forward. In the survey, there were three options for ‘married’, they were: 1=Never married, 2=Currently married, 3=Previously married.

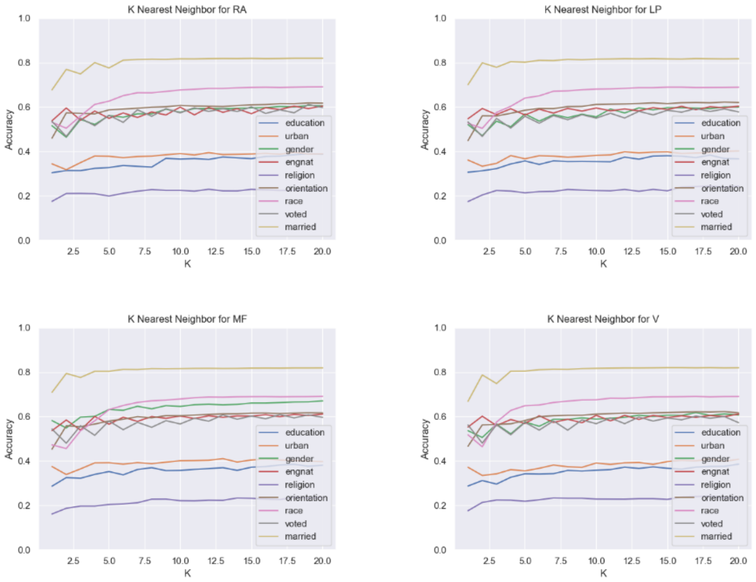


Figure 3. Top-Left – KNN for ‘RA’. Top-Right – KNN for ‘LP’. Bottom-Left – KNN for ‘MF’. Bottom-Right – KNN for ‘V’.

In order to get the most accurate model, we need to pick the most accurate K value. We can do that by calculating the error rate for a range of K’s, plotting, and using the elbow method to pick the best one.

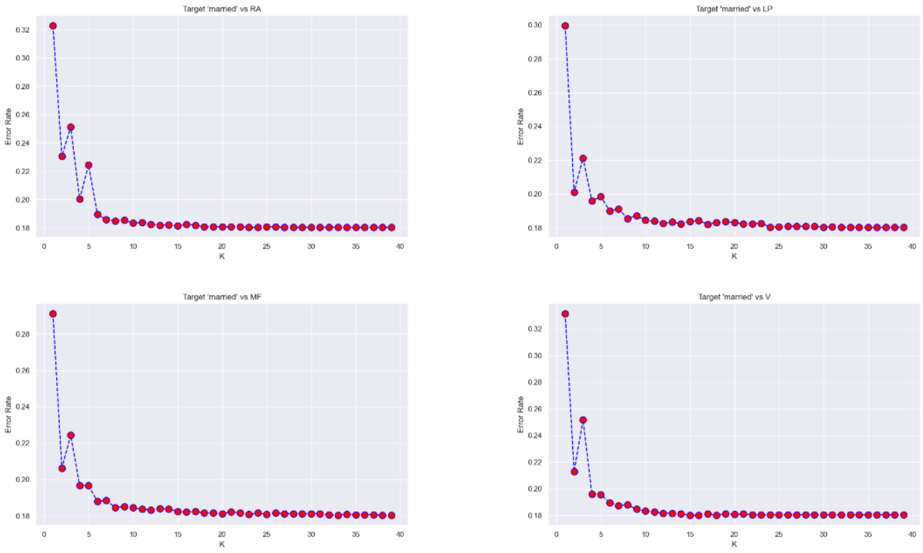
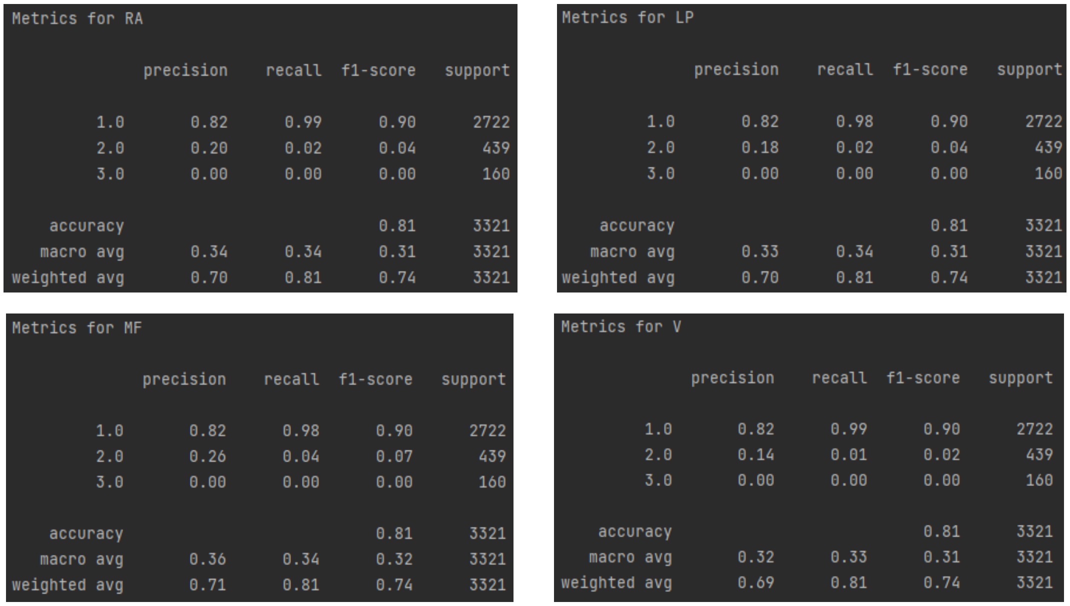


Figure 4. Error rates for many K values where the target is 'married'.

As shown in the 4 graphs above, the elbow is at K = 6 for 3 of the 4, so that will be the number to train the model.



```
Metrics for RA

              precision    recall  f1-score   support

         1.0       0.82      0.99      0.90      2722
         2.0       0.20      0.02      0.04       439
         3.0       0.00      0.00      0.00       160

    accuracy                           0.81      3321
   macro avg       0.34      0.34      0.31      3321
weighted avg       0.70      0.81      0.74      3321
```

```
Metrics for LP

              precision    recall  f1-score   support

         1.0       0.82      0.98      0.90      2722
         2.0       0.18      0.02      0.04       439
         3.0       0.00      0.00      0.00       160

    accuracy                           0.81      3321
   macro avg       0.33      0.34      0.31      3321
weighted avg       0.70      0.81      0.74      3321
```

```
Metrics for MF

              precision    recall  f1-score   support

         1.0       0.82      0.98      0.90      2722
         2.0       0.26      0.04      0.07       439
         3.0       0.00      0.00      0.00       160

    accuracy                           0.81      3321
   macro avg       0.36      0.34      0.32      3321
weighted avg       0.71      0.81      0.74      3321
```

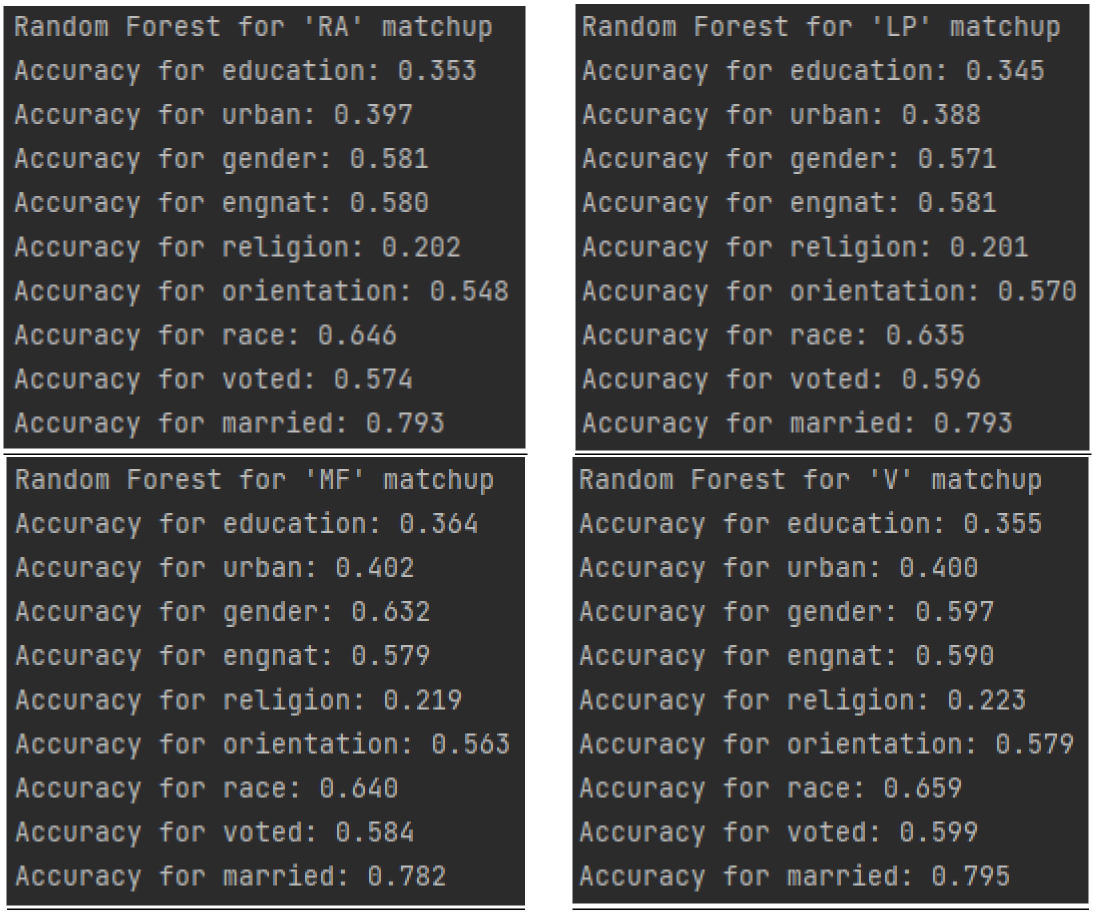
```
Metrics for V

              precision    recall  f1-score   support

         1.0       0.82      0.99      0.90      2722
         2.0       0.14      0.01      0.02       439
         3.0       0.00      0.00      0.00       160

    accuracy                           0.81      3321
   macro avg       0.32      0.33      0.31      3321
weighted avg       0.69      0.81      0.74      3321
```

Figure 5. Metrics for RA, LP, MF, and V matchups for K = 6.

This model performs fairly well and as we can see above, the accuracy is 0.81 for each matchup. What we can also see is that the model performs terribly (0% correct) in predicting a value of 3, which is 'Previously married'. The model doesn't predict a value of 2 correctly much either. This can probably be attributed to the distribution of responses which are skewed heavily to response 1.

RF

Below are these results from running each matchup separately through an RF classifier. Using a test size of 20%, we use every painting from the matchup to predict various targets. Again, as with KNN, there are targets that perform well and some that perform poorly. Religion, Education, and Urban are again bringing up the rear while Married performs the best again.



```
Random Forest for 'RA' matchup
Accuracy for education: 0.353
Accuracy for urban: 0.397
Accuracy for gender: 0.581
Accuracy for engnat: 0.580
Accuracy for religion: 0.202
Accuracy for orientation: 0.548
Accuracy for race: 0.646
Accuracy for voted: 0.574
Accuracy for married: 0.793
```

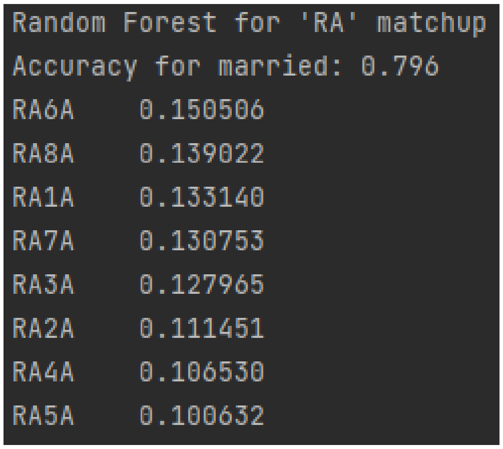
```
Random Forest for 'LP' matchup
Accuracy for education: 0.345
Accuracy for urban: 0.388
Accuracy for gender: 0.571
Accuracy for engnat: 0.581
Accuracy for religion: 0.201
Accuracy for orientation: 0.570
Accuracy for race: 0.635
Accuracy for voted: 0.596
Accuracy for married: 0.793
```

```
Random Forest for 'MF' matchup
Accuracy for education: 0.364
Accuracy for urban: 0.402
Accuracy for gender: 0.632
Accuracy for engnat: 0.579
Accuracy for religion: 0.219
Accuracy for orientation: 0.563
Accuracy for race: 0.640
Accuracy for voted: 0.584
Accuracy for married: 0.782
```

```
Random Forest for 'V' matchup
Accuracy for education: 0.355
Accuracy for urban: 0.400
Accuracy for gender: 0.597
Accuracy for engnat: 0.590
Accuracy for religion: 0.223
Accuracy for orientation: 0.579
Accuracy for race: 0.659
Accuracy for voted: 0.599
Accuracy for married: 0.795
```

Figure 6. Random Forest accuracy for all Matchups.

The accuracy for all targets seems to be near identical for every matchup. A deeper look will be taken at the RA matchup to see if we can improve the accuracy. An advantage of RFs is being able to take a look at the relative feature importance. To calculate this score, RF uses Gini importance or mean decrease in impurity to calculate the importance of each feature. This helps in identifying and selecting the most contributing features for the classifier. The least important feature(s) can be removed to increase the accuracy of the model. Shown below are the Gini importance values for the RA matchup where the target is Married, as this is again the most accurate feature across all four matchups. Ranked from most important to least, it can be seen that the painting RA6A is the most important in the prediction, while RA5A is the least important feature. The range in values is not that high, this indicates that each feature is of almost equal importance in the prediction of the RF.

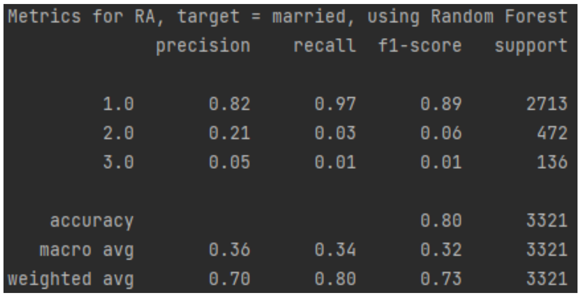


```
Random Forest for 'RA' matchup
Accuracy for married: 0.796
RA6A    0.150506
RA8A    0.139022
RA1A    0.133140
RA7A    0.130753
RA3A    0.127965
RA2A    0.111451
RA4A    0.106530
RA5A    0.100632
```

Figure 7. Gini Importance values for RA matchup with target = Married.

When the worst feature, RA5A, is removed and a new model is created with only 7 features, the accuracy actually decreases slightly to 0.792, and the new worst feature is RA2A with a score of 0.11. Removing the worst feature again and re-training the model leads to better results with an accuracy of 0.805. It seems that with all of the feature importance values being in such a small range, removing the worst doesn't really seem to improve the model that significantly (or it could decrease the accuracy). If the most important feature had a feature importance value that was several times higher than the least important, removing the least important feature would have a much better impact.

Overall, the accuracy of this RF model is fairly good. Just like the KNN model, the accuracy is this high because the data is so skewed towards a marriage response of 'never married'. However, when looking at the overall metrics of the RF and not just the accuracy, it can be seen that it has outperformed the KNN model. This RF model outperformed the KNN's f-1 score with respect to response '3' by 0.01, and response '2' by 0.02 for the RA matchup.



```
Metrics for RA, target = married, using Random Forest
              precision    recall  f1-score   support

         1.0       0.82      0.97      0.89      2713
         2.0       0.21      0.03      0.06       472
         3.0       0.05      0.01      0.01       136

    accuracy                           0.80      3321
   macro avg       0.36      0.34      0.32      3321
weighted avg       0.70      0.80      0.73      3321
```

Figure 8. Metrics for the RA matchup using the RF model.

4. Conclusion

Despite the KNN model’s ability to accurately predict ‘never married’, the data was heavily skewed towards that response. Even a terribly fit model, such as this one, would accurately predict ‘never married’ because of the heavily skewed data. The same applies to the RF modeling done in this project. In order to create better models, these classification analyses would need to be done on data that is more evenly distributed, especially among the marriage status responses. Additionally, it is my recommendation that KNN and RF classifications should not be used in order to analyze Likert scale data with 4+ response options. I believe the Likert scale was too large to predict within my models due to the fact that you have two responses for liking (or disliking) the paintings. I believe that a Likert scale response on a 3-point scale would be more accurate. To wrap up, I recommend that future KNN and RF classification modeling projects to be done on datasets with more binary or dichotomous outcomes.

References

1 *Impressionism*. (2017). Harvard.edu. https://scholar.harvard.edu/mourad/section

2 *Realism In Art: What It Means & Why It's Valuable For Practicing Artists*. (2018, July 28). Concept Art Empire. https://conceptartempire.com/realism/

3 Blumberg, N. (2019). Landscape painting | art. In *Encyclopædia Britannica*. https://www.britannica.com/art/landscape-painting

4 Tate. (2017). *Surrealism – Art Term | Tate*. Tate. https://www.tate.org.uk/art/art-terms/s/surrealism

5 Harshi, H. (2021, July 26). *Understanding K-Nearest Neighbour Algorithm in Detail*. Analytics Vidhya. https://medium.com/analytics-vidhya/understanding-k-nearest-neighbour-algorithm-in-detail-fc9649c1d196

6 Avinash Navlani. (2018). *Random Forests Classifiers in Python*. DataCamp Community. https://www.datacamp.com/community/tutorials/random-forests-classifier-python