The Erdős Data Science Bootcamp Fall 2024 Executive Summary

Team Members: Greyson Meyer, Sun Lee, Dawson Kinsman, Oulin Yu, Botan Cevik

Github: https://github.com/BotanCevik2/Project-Leonardo

Overview:

Project Leonardo, (or *Using Image Recognition to Find Similar Art)*, aims to develop an innovative recommendation system that identifies and suggests visually similar paintings based on color and composition analysis. Leveraging advanced image recognition techniques, we evaluate how well statistical techniques, specifically k-means clustering, capture and represent dominant visual features of artwork.

Objective:

The project aims to extract and store meaningful visual features (color and composition) from a dataset of images using k-means clustering. By applying k-means with varying numbers of clusters (k), we analyze how effectively the algorithm identifies key colors and structural features in the images. For color, we cluster pixel values, while for composition, we analyze contours and shapes in the image. Performance is assessed by measuring the distance between the cluster centroids of the input image and those in the dataset. Based on these features, we recommend visually similar images and visualize the top matches for both color and composition, helping to assess the efficacy of the clustering algorithm in art recommendation.

Stakeholders:

Art curators, people looking for similar art pieces, scholars.

Dataset:

We used the resticted version of OmniArt dataset consisting of over 15 online artwork collections and user generated art uploaded on the internet.

Data Processing: The project involves preparing a large dataset of image URLs and associated metadata for analysis and recommendations. The data preprocessing steps include:

- **Data Cleaning:** Removing entries with missing artist names, broken image URLs, or irrelevant categories (e.g., non-painting-related entries) to ensure the dataset focuses solely on paintings and related artwork.
- **Image Downloading and Resizing:** Downloading the valid images from their URLs and resizing them to a consistent dimension for uniformity in feature extraction and analysis.
- **Efficient Storage:** Organizing and storing the cleaned dataset, including resized images and their metadata, in Parquet files. This format allows for efficient retrieval and processing of large-scale data during clustering and similarity evaluation.

Performance Analysis:

- Evaluating recommendations based on their alignment with input image features.
- Silhouette scores serve as a quantitative metric to validate the clustering quality for both color and composition.

Finally, we've written a web app Project Leonardo .

Future Iterations:

-Improved Feature Extraction:

Incorporate additional visual features such as texture, edge density, or artistic styles using deep learning-based techniques (e.g., convolutional neural networks).

- Dynamic Clustering Method:

Investigate clustering with varying *k* based on individual image complexity instead of a fixed *k* value.

-User Interface:

Develop a user-friendly interface to allow interactive exploration of recommendations.