Term Project Phase I

POC: Michael Mogilevsky (*mm3201*)
Adarsh Narayanan (*ahn38*)
Hazem Zaky (*hgz5*)

Team Name: The Sigmoid Squad

November 2024

1 Introduction

In this phase of the term project, we have selected four distinct datasets as per the instructions for a three person group and plan to explore various machine learning tasks through these datasets, covering classification, regression, and clustering. Each dataset is summarized below, along with the associated machine learning task and source link.

2 Datasets and Machine Learning Tasks

2.1 Online Gaming Anxiety Data

Description: The *Online Gaming Anxiety Data* dataset, collated by Marian Sauter and Dejan Draschkow, examines the association between online gaming behaviors and psychological factors such as anxiety, life satisfaction, and social phobia. It includes 13,500 data samples, providing demographic details (age, gender, education level), gaming habits (weekly gaming hours, preferred platforms), and assessments using standardized scales like the Generalized Anxiety Disorder (GAD) scale, Satisfaction with Life Scale (SWLS), and Social Phobia Inventory (SPIN). This dataset facilitates analysis of correlations between gaming activities and mental health indicators.

Source: https://www.kaggle.com/datasets/divyansh22/online-gaming-anxiety-data

Machine Learning Task: Classification

Motivation: We are all gamers who enjoy video games, and coming from the COVID-19 era, where online gaming was a significant source of social interaction, it is interesting to explore the potential anxiety caused by online gaming and its associated social interactions.

Objective: Develop a model to predict the likelihood of a participant experiencing anxiety based on features like demographic details, gaming habits, and psychological assessment scores.

2.2 Financial Data

Description: The *Gold Price Prediction* dataset provides historical financial data, including daily prices of various financial instruments such as stocks, commodities, and indices. It spans from 2010 to 2024, comprising approximately 3904 data samples. Each data sample includes features such as opening and closing prices, highest and lowest prices of the day, trading volume, and other relevant financial indicators. The time-series nature of this data allows for in-depth analysis of trends and patterns over time, making it suitable for forecasting future prices or returns.

Source: https://www.kaggle.com/datasets/franciscogcc/financial-data

Machine Learning Task: Regression

Motivation: It is fascinating to observe correlations between major indices like the S&P 500 and Nasdaq, as well as the behavior of individual stocks. This dataset provides an opportunity to explore financial market trends and predictions.

Objective: Utilize historical financial data to predict future prices or returns of specific financial instruments, aiding in investment decisions and risk management.

2.3 Small Image Dataset for Unsupervised Clustering

Description: The "Small Image Dataset for Unsupervised Clustering" contains 80 images divided amoungst the following distribution: dogs (10), cats (10), family (20), alone (20), and food (20). Each image is unlabeled, requiring analysis based solely on its visual characteristics.

Source: https://www.kaggle.com/datasets/heavensky/image-dataset-for-unsupervised-clustering

Machine Learning Task: Clustering

Motivation: While foundational datasets like MNIST, STL-10, CIFAR-10, and ImageNet are widely used, we specifically chose an image dataset tailored for unsupervised clustering. Distance metrics between image pixel such as the known Wasserstien distance (also known as Earth Mover's Distance) could be interesting when used in analysis for machine learning to measure dissimilarity between two probability distributions, and this dataset meets our requirements for experimenting with clustering methods.

Objective: Implement unsupervised learning techniques to group similar images based on visual features, facilitating tasks like image organization, pattern recognition, and anomaly detection.

2.4 Coronavirus Tweets NLP - Text Classification

Description: The "Coronavirus Tweets NLP - Text Classification" dataset contains 3,798 tweets during the COVID-19 pandemic. Each tweet is annotated with one of five sentiment labels: extremely positive, positive, neutral, negative, or extremely negative. The dataset includes the text of each tweet, the user's location, and the date it was posted, enabling detailed analysis of public sentiment and trends during the early phase of the pandemic.

Source: https://www.kaggle.com/datasets/datatattle/covid-19-nlp-text-classification

Machine Learning Task: Classification

Motivation: Twitter posts provide a great source of text data, as they reflect opinions from users worldwide. This dataset focuses on tweets about the coronavirus during a time of reduced social interaction. Analyzing sentiments during this period offers insights into how the pandemic affected public opinions, especially regarding negative sentiments.

Objective: Develop a model to classify the sentiment expressed in tweets, aiding in sentiment analysis, trend detection, and public opinion monitoring during health crises.