Dataset Overview

For this project, I will be using a Steam games data set. It consists of 4 csv files, two being from March 2024 with one being cleaned for duplicates/playtest games, and the other two being from May 2025, again, one cleaned and one not. I will likely be using the uncleaned version of the May 2025 file in my project. In this file, there are 47 columns, and over 80,000 entries per column. Many of the columns do not suit what I would like to do with the data set, so they will likely be scrubbed when training the algorithms. I chose this data set because I love video games and I thought the sheer amount of data in the set would be a challenge to clean and go through, which I am excited to do.

Link: <https://www.kaggle.com/datasets/artermiloff/steam-games-dataset/data>

Classification/Regression?

For my task, I want to make a recommendation to a user to see if they will like a game or not based on various other games played. When training my algorithms, this would look like: based on certain features, predict whether or not a user will like the next game shown. This would be a problem of binary classification, the binary factors being yes/no.

Problem Statement

Given reviews, description, Metacritic score, price, developers, and many other features of multiple games, can we predict if the user will like the next game shown.

**Project discussion**

Using positive and negative reviews as the yes/no column. Use at least 10 features to generate the prediction

Decide a threshold of positive/negative reviews (this game has a much higher positive review amount than negative) and train the algorithm based on that

Use scikit-learn to give examples of how to implement the algorithms in code

Can use logistic regression

Try to find out if some papers have used this data or not

If a paper is found that is more complex than this one, you can reference it but you can say that this is a simpler version

Create a recommendation column for yes and no from positive and negative columns

If (positive reviews > (threshold, i.e. 30,000))

If (positive reviews > negative reviews)

Recommendation = yes

Else if (positive reviews < negative reviews)

Recommendation = no

Must add recommendation column to the csv file

Real python tutorials for python tutorials with virtual environments and practices for codes

Pandas tutorial on geeks for geeks