Machine Learning Journal

**Name:** Alfredo Garza  
**Lab Title:** Video Game Sales Prediction  
**Date:** July 24, 2025

Objective

To develop a classification model that predicts whether a video game will be a commercial “hit” based on metadata and global/regional sales performance. A game qualifies as a "hit" if it surpasses 1 million units sold globally.

Dataset Overview

* **Source:** Kaggle (Gregorut’s Video Game Sales Dataset)
* **Size:** 16,598 records
* **Features Used:** Platform, Genre, Publisher, Year, NA\_Sales, EU\_Sales, JP\_Sales, Other\_Sales, Global\_Sales
* **Target Variable:** Hit — binary indicator (1 = sales > 1M units)

Data Preprocessing

* Imputed missing Year values using median.
* Encoded categorical columns with LabelEncoder.
* Dropped Rank and Name columns as non-predictive.
* Engineered a Total\_Regional\_Sales feature to validate consistency with Global\_Sales.

Data Exploration

* Only **~12.4%** of games were classified as "hits" — major class imbalance.
* Visualized distribution of Global\_Sales and target variable.
* Confirmed correlation between Total\_Regional\_Sales and Global\_Sales.

Model Training

* Model: DecisionTreeClassifier (default hyperparameters)
* Dataset split: 80% train, 20% test
* Achieved perfect scores across all metrics:
  + **Accuracy:** 1.000
  + **Precision:** 1.000
  + **Recall:** 1.000
  + **F1 Score:** 1.000
  + **ROC AUC:** 1.000
* Visualized confusion matrix — no false positives or negatives

Reflections

The perfect metrics strongly suggest overfitting. This model likely memorized training data rather than learning generalizable patterns. This is a key learning moment — accuracy isn’t everything.

Next Steps

* Tune hyperparameters using GridSearchCV (e.g. max\_depth, min\_samples\_split, min\_samples\_leaf)
* Explore pruning or switch to ensemble models (e.g. RandomForest, XGBoost)
* Apply k-fold cross-validation to confirm model reliability
* Review feature importance rankings to refine inputs

Would you like to build this into a formal report with visual sections, or integrate it into your Golden Verba write-up to show semantic analysis potential? I could even help you turn it into a blog or portfolio page.