**Project Report: Linear Regression and KNN for Global Sales Prediction**

**1. Introduction**

This project aims to compare the performance of two machine learning models—Linear Regression and K-Nearest Neighbors (KNN)—in predicting global sales in video games. The dataset consists of various features, including game release year, platform, genre, publisher, and global sales figures.

**Objectives:**

* To apply Linear Regression and KNN for predicting global sales.
* To compare model performance based on various metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), R-squared (R2), and ROC-AUC curves.
* To visualize the error distribution, actual vs. predicted values, and key performance indicators.

**2. Dataset Overview**

The dataset used for this project is sourced from a CSV file containing game sales data. The relevant columns considered for the analysis are:

* **Year**: The release year of the game.
* **Platform**: The gaming platform the game was released on.
* **Genre**: The genre of the game.
* **Publisher**: The game's publisher.
* **Global\_Sales**: The total global sales of the game (target variable).

**Preprocessing Steps:**

* **Missing Data Handling**: Missing numeric data for "Year" and "Global\_Sales" were imputed using the mean strategy. Missing categorical data for "Platform," "Genre," and "Publisher" were filled with the mode (most frequent value).
* **Feature Engineering**: Categorical variables were one-hot encoded, and numeric variables were scaled using the StandardScaler.
* **Outlier Removal**: Outliers were removed by applying Z-scores for numerical features.
* **Data Splitting**: The dataset was split into training and testing sets (80%-20%).

**3. Model Implementation**

**3.1 Linear Regression**

**Model Overview:**

Linear Regression is used to predict the target variable, **Global\_Sales**, based on the other features in the dataset. The model assumes a linear relationship between the input features and the target.

**Training the Model:**

The Linear Regression model was trained using the training set. The coefficients were learned using ordinary least squares (OLS).

**Evaluation Metrics:**

The model's performance was evaluated using the following metrics:

* **Mean Squared Error (MSE)**: 0.1636
* **Mean Absolute Error (MAE)**: 0.2381
* **R-squared (R2)**: 0.1249

**Graph: Loss Curve (Manual Linear Regression)**

This graph shows the decrease in training loss (MSE) over 1000 epochs, indicating how the model learns to minimize the error.

A blue line on a white graph

Description automatically generated**3.2 K-Nearest Neighbors (KNN)**

**Model Overview:**

The KNN algorithm is a non-parametric method used for regression, where the prediction is made based on the average of the closest neighbors in the feature space.

**Training the Model:**

The KNN model was trained using the scaled features with 5 neighbors.

**Evaluation Metrics:**

The KNN model's performance was evaluated using the following metrics:

* **Mean Squared Error (MSE)**: 0.1684
* **Mean Absolute Error (MAE)**: 0.2356
* **R-squared (R2)**: 0.0994

**4. Model Comparison**

**4.1 Error Distribution**

A histogram was plotted to compare the error distributions of Linear Regression and KNN. The errors were calculated as the difference between the actual and predicted global sales.

* [A graph of a error distribution

  Description automatically generated]

**4.2 Accuracy Metrics**

The R-squared values for both models were calculated to assess their fit to the data.

* **Linear Regression R2 Score**: 0.1249
* **KNN R2 Score**: 0.0994

Both models show relatively low R-squared values, indicating that the models do not explain much of the variance in the target variable.

**4.3 Precision and Recall (Proxy using Regression Errors)**

* **Mean Absolute Error (MAE)**:
  + Linear Regression: 0.2381
  + KNN: 0.2356
* **Root Mean Squared Error (RMSE)**:
* A graph of blue bars with black text

  Description automatically generated
  + Linear Regression: 0.4045
  + KNN: 0.4104

**4.4 ROC and AUC (Pseudo Classification by Thresholding)**

For the ROC curve, predictions were scaled to [0, 1], and the models were evaluated as a binary classification task by thresholding the target variable at its mean.

A graph of a curve

Description automatically generated with medium confidence**5. Results and Discussion**

The models' evaluation shows that neither Linear Regression nor KNN performed exceptionally well. The R-squared values suggest that both models have a limited ability to predict global sales based on the given features. However, KNN had a slightly better performance in terms of MSE and MAE, though it still did not capture a significant portion of the variance in the data.

Both models can benefit from additional feature engineering, hyperparameter tuning, or perhaps using more sophisticated algorithms (e.g., decision trees, random forests) to improve predictive power.

**6. Conclusion**

In this project, we compared two machine learning models, Linear Regression and KNN, for predicting global video game sales. While the models showed decent evaluation scores, further improvements can be made with additional model optimization. More advanced techniques like feature selection, parameter tuning, and cross-validation could enhance the model's predictive performance.