ACM 476 TERM PROJECT REPORT

Title: Term Project
Submitted By: Batuhan Kaya – Yasin Aydın
Submitted To: Ayşe Başar
Date: 3.06.2025
Section: 1

1 Purpose of Project:
The main goal of this project is to identify the key factors that affect video game sales performance in the gaming industry and to develop predictive models based on this data. In this context, four main objectives have been defined:
• Identify critical variables that impact game success
• Compare feature selection and dimensionality reduction methods
• Develop models that classify games as successful/unsuccessful
• Build regression models to predict global sales volume

1.2 Dataset Description
• Source: Video Game Sales as of December 22, 2016 (Kaggle)
• Sample: 2,000 games randomly selected from 16,000+ games
• Feature Types:
  - Categorical: Platform, Genre, Publisher, Developer, Rating
  - Numerical: Regional sales (NA, EU, JP, Other), Critic/User scores
  - Target Variables: Global sales volume and success class (binary)

2. Methodology

2.1 Data Preprocessing
• Missing values: Filled with median and mode
• Categorical data: One-hot encoding
• Numerical data: Standardization
• Feature engineering: Success label, sales ratios, interaction variables

2.2 Feature Selection & PCA
• Methods used: ANOVA, Mutual Info, Random Forest, RFE
• The first 3 components from PCA explained 85% of the variance
• Feature selection generally outperformed PCA

2.3 Clustering
• Hierarchical clustering (Ward + Euclidean)
• Three clusters determined using the Elbow method
• Clusters were differentiated by platform, genre, and sales level

2.4 Classification
• Models: Random Forest, Logistic Regression, SVM
• Evaluation: Accuracy, Precision, Recall, F1-score
• Best result: Random Forest (78% accuracy)

2.5 Regression
• Models: KNN, Linear, Random Forest
• Best performance: KNN (R^2 = 0.72)
• Especially accurate in mid-level sales predictions

3. Results

3.1 Prominent Features
• NA_Sales, EU_Sales, Year_of_Release, and Platform
• ANOVA and RF provided the highest accuracy (0.78)

3.2 Clustering
• Cluster 1: AAA games
• Cluster 2: Mid-level games
• Cluster 3: Niche, low-sales games

4. Challenges and Solutions
• Missing data: Filled with median/mode
• Class imbalance: Balanced with SMOTE
• Outliers: Handled with robust scaling
• Computation time: Optimized with parallel processing

STEP	PROCEDURE	OBSERVATIONS
1	Data Upload	Use of UTF-8 Encoding
2	Missing Data Analysis	Median Imputation and Alternative Analysis Methods
3	Data Cleaning	Standardized platform categories
4	Feature Transformation	One-hot encoding and merging of categories
5	Feature Selection	Use of Recursive Feature Elimination (RFE)
6	Classification	Sample balancing with SMOTE
7	Regression	Robust scaling ve outlier handling
8	Clustering	Elbow method and silhouette analysis
9	Computation Time	Parallel processing and batch processing
10	Memory Usage	Optimization of data types
11	Model Training	Early stopping ve hyperparameter tuning
12	Code Efficiency	Functional programming and modular structure

CONCLUSIONS AND DISCUSSION Throughout this project, we gained a great deal of knowledge both technically and analytically. First and foremost, we learned that data mining processes are not just about writing code; it is a multifaceted journey that begins with properly cleaning, transforming, and processing the data.

-We learned how to deal with missing data. Not every missing value should be deleted; appropriate imputation methods can preserve data quality.

-We saw how impactful feature selection can be on model performance. Removing unnecessary features results in faster and more accurate models.

-By comparing different classification and regression models, we learned the strengths and weaknesses of each in various scenarios.

-Through clustering analysis, we discovered how revealing hidden structures in the data can be a powerful tool for decision support.
-We directly experienced the challenges of working with raw and complex real-world data.
-We realized that generating insights for business is not only about performing correct analysis but also effectively communicating those insights.
-While working as a team, we experienced the importance of task distribution, version control, and documentation of code.
-Most importantly, we understood that data offers us not just numbers but meaning—and as long as we can extract that meaning, we can create real value from data.
This project served not only as a course assignment but also laid a strong foundation for our future endeavors in the field of data science. With every analysis, we asked ourselves: "how can we do this better?"