# Advanced Data Analytics

## Exercise project 1

### 1. Data Overview

**Classification Dataset**: Mushroom

- Dataset: The Mushroom Classification dataset was used to predict whether a mushroom is edible or poisonous based on its features.
- Features Analyzed: Attributes such as cap shape, color, gill size, odor, and more.
- Target Variable: Edibility (edible or poisonous).

**Regression Dataset**: VGChartz

- Dataset: The VGChartz dataset was used for predicting numerical values related to video game sales.
- Features Analyzed: Game titles, platform, year of release, user ratings, and sales figures.
- Target Variable: Total sales in millions.

## 2. Data Preprocessing and Challenges

### Imbalance Handling

Both datasets initially exhibited class or distribution imbalances:

- **Classification Imbalance**: The Mushroom dataset had a relatively balanced distribution between edible and poisonous, but certain features had skewed distributions.
- **Regression Imbalance**: The VGChartz data had a heavy skew towards lower sales figures.

### Outlier Detection and Handling

I applied **Isolation Forest** to identify and mitigate the impact of outliers:

- **Effectiveness:** The technique isolated anomalies effectively, improving the data quality for more accurate predictions.

### Noise Management

Noise was managed in both datasets to reduce variability while preserving important information:

- **Mushroom Dataset**: Managed through feature encoding and data cleaning.
- **VGChartz Dataset**: Cleaned user ratings and other variables to handle noise.

## 3. Model Building

**Classification Model**: Used models such as decision trees and random forests after preprocessing to classify mushrooms, with improved performance post-cleaning.

**Regression Model**: Applied models such as linear regression and ensemble-based algorithms to the VGChartz dataset, benefiting from outlier management.

## 4. Results and Observations

**Classification Results:**

- **Accuracy**: Improved with the removal of outliers, making the model better at distinguishing between edible and poisonous mushrooms.
- **Precision/Recall**: Metrics indicated a balanced classification with reduced false positives/negatives.

**Regression Results:**

- **R-Squared Value**: Increased significantly after outlier management, indicating a stronger model fit.
- **Mean Squared Error**: Lowered, showing improved predictive performance.

## Conclusion

Using Isolation Forest for outlier management effectively refined data quality for both the Mushroom and VGChartz datasets. This led to improved model performance and more reliable predictive analysis, addressing initial challenges like noise and imbalances.