**INST414 Final Project**

Team Members: Jim Chen, Christopher Aguila, Nicholas Koy, Gerald Reyes, Kaveh Vakili

**Introduction**

Our source of data is the dataset “Video Game Sales”; it is a dataset on Kaggle and it can be found through the following link: <https://www.kaggle.com/datasets/gregorut/videogamesales>

This dataset contains information about video game sale amounts around the world.

The dataset has 16,598 rows and 11 columns; the columns are as follows: “Rank”, “Name”, “Platform”, “Year”, “Genre”, “Publisher”, “NA\_Sales”, “EU\_Sales”, “JP\_Sales”, “Other\_Sales”, “Global\_Sales”.

Our target column is the “Genre” column while the other columns are our features columns.

There are 12 genres a video game can be classified into: “Sports”, “Platform”, “Racing”, “Role-Playing”, “Puzzle”, “Misc”, “Shooter”, “Simulation”, “Action”, “Fighting”, “Adventure”, “Strategy”

**Questions**

Questions we seek to answer are:

* With Logistic Regression:
  + How accurately can a video game’s genre be predicted when the game’s rank, global sales amount, platform, release date year, and publisher are known?
* With Decision Tree
  + How accurately can a video game's genre be predicted if the game's rank, name, release date year, publisher, and global sales amount are known?
  + What columns will be used to create branches in our decision tree model?
* With XGBoost
  + How accurately can a XGBoost model predict a video game's genre when the video game's rank, global sales amount, platform, release date year, and publisher are known?
  + Which feature is most important in predicting a video game’s genre?
  + Which feature is least important in predicting a video game’s genre?
* With Neural Network (Keras)
  + What is the highest accuracy a keras neural network can reach when predicting a video game's genre based on the game's rank, name, platform, release year, publisher, and global sales amount?
  + What is the lowest accuracy a keras neural network can reach when predicting a video game's genre based on the game's rank, name, platform, release year, publisher, and global sales amount?
  + Is there a major change in accuracy between epochs?
* With Neural Network (PyTorch)
  + What is the highest accuracy a PyTorch neural network can reach when classifying a video game's genre based on the game's rank, name, platform, release year, publisher, and global sales amount?
  + What is the lowest accuracy a PyTorch neural network can reach when classifying a video game's genre based on the game's rank, name, platform, release year, publisher, and global sales amount?
  + Is there a major change in accuracy between epochs?

**Findings**

* Logistic Regression:
  + A video game’s genre cannot be accurately predicted with just a video game's rank, global sales amount, platform, release date year, and publisher
  + Our logistic regression model was overall 35% accurate.
  + One thing to note though is that our logistic regression model was able to best predict the Adventure genre; it predicted this genre correctly 51% of the time. The relatively high accuracy in this genre means this genre does somewhat have a correlation with a video game's rank, global sales amount, platform, release date, year, and publisher.
* Decision Tree
  + A video game’s genre cannot be accurately predicted with just a video game's rank, name, release date year, publisher, and global sales amount
  + Our decision tree model was over all ~21% accurate
  + The columns our decision tree model used to create branches are “Year”, “Platform”, and “Publisher”
* XGBoost
  + A video game’s genre cannot be accurately predicted with just a video game's rank, global sales amount, platform, release date year, and publisher
  + Our XGBoost model was overall 35% accurate.
  + The most important feature in predicting a video game's genre in our XGBoost model is a video game’s rank.
  + All other features in our dataset appear to have relatively low importance; this potentially due to pd.get\_dummies breaking these features up into multiple columns
* Neural Network (Keras)
  + With a keras neural network, we cannot accurately predict a video game's genre when the game's rank, name, platform, release year, publisher, and global sales amount are known.
  + Our keras neural network model's accuracy peaked at epoch 4 with an accuracy of around 19.94%
  + The lowest accuracy our keras neural network model had was at epoch 1 with an accuracy of 14.34%
  + Our keras neural network model's accuracy did not change a lot as the epochs progressed; the accuracy increased by about 5% over 20 epochs.
* Neural Network (PyTorch)
  + A video game’s genre prediction model using PyTorch neural networks performs worse than its Keras counterpart when using the game’s rank, name, platform, release year, publisher, and global sales as features.
  + Using Pytorch, the highest accuracy our multiclass classification model reached was about 15.56%; this was reached at epoch 20
  + Using Pytorch, the lowest accuracy our multiclass classification model reached was about 11.81%; this was reached at around epoch 3
  + Overall, from epoch 1 to epoch 20, the accuracy of our PyTorch neural network model increased by around 4% primarily fluctuating between ~14% and 15%

**Challenges**

The primary challenge was that most examples involved binary classification. It took a significant amount of time to adapt these models for multiclass classifcation.

* Decision Tree
  + The in-class examples of Decision Trees dealt with binary classification, not classification with 12 different results
  + As such, we had to use outside resources to learn how to adapt our decision tree to classifying more than two values
* XGBoost
  + The importance of features may be biased because while the rank column was never broken, other columns, like Year and Publisher, were broken into many different columns through the use of pd.get\_dummies()
* Neural Network (PyTorch)
  + Hard to find resources and examples for multiclass classification.
  + Error with tensor sizing for loss criterion function, and had to spend time debugging out the error which turned out to be turning a 2d tensor into a 1d tensor.

**Comparisons**

When it comes to predicting a video game’s genre, whcn all 12 genres are included, our models had the following accuracy:

* Logistic Regression - 35%
* Decision Tree - 21%
* XGBoost - 35%
* Neural Network (Keras) - 20%
* Neural Network (PyTorch) - 15%

From these accuracies, we can see that a Logistic Regression model and XGBoost model are equally best at predicting a video game’s genre, when all 12 genres are included; they both had an accuracy of about 35%.

Our least accurate model was the PyTorch neural network model; it had an accuracy of about 15%.

**Conclusion**

When predicting a video game’s genre and its rank, name, platform, release date year, and global sales amount are known, it is best to use a Logistic Regression model or a XGBoost model.