

CS-7641
Assignment 1- Supervised Learning
Nayeem Aquib

1 INTRODUCTION

In this report, 5 different learning algorithms- Decision trees, Neural networks, Boosting, Support Vector Machines, and k-nearest neighbors, will be used to train and test on 2 different datasets. Both of these datasets were acquired from Kaggle. One is a dataset of videogame sales across different platforms and different regions. Another is a dataset of netflix movies and tv shows across different genres and countries all over the world.

2 CLASSIFICATION PROBLEMS

2.1 Dataset - Video Game Sales

2.1.1 Description 1

Using a dataset and a learning algorithm, we will classify whether a video game will be a "success" or a "failure" based on its global sales.

2.1.2 Why is it Interesting

1. I am curious to see which games come up on top and if it matches the hype.
2. To check if based on this data, my model can predict a nintendo games success such as Super Mario Bros: Wonder in future.
3. From a more practical perspective with a business motivation, a solution to this problem could provide some guidance on risk assessment and market strategy for a game publisher like Nintendo before releasing a game.

2.2.1 Description 2

Using a dataset and a learning algorithm, we will predict whether a video game belongs to a specific genre (e.g., "Action") or not based on its features.

2.1.2 Why is it Interesting

1. It is important for curation and licensing purposes.
2. Based on the demographics and region the game is being released, it could be banned in some places.

3 THE LEARNING ALGORITHMS ANALYSIS

The SVM performed quite well on the both training and testing dataset. The accuracy was at 0.99 for both training and testing dataset. The accuracy improved substantially once the number of data points increased to more than 7900.

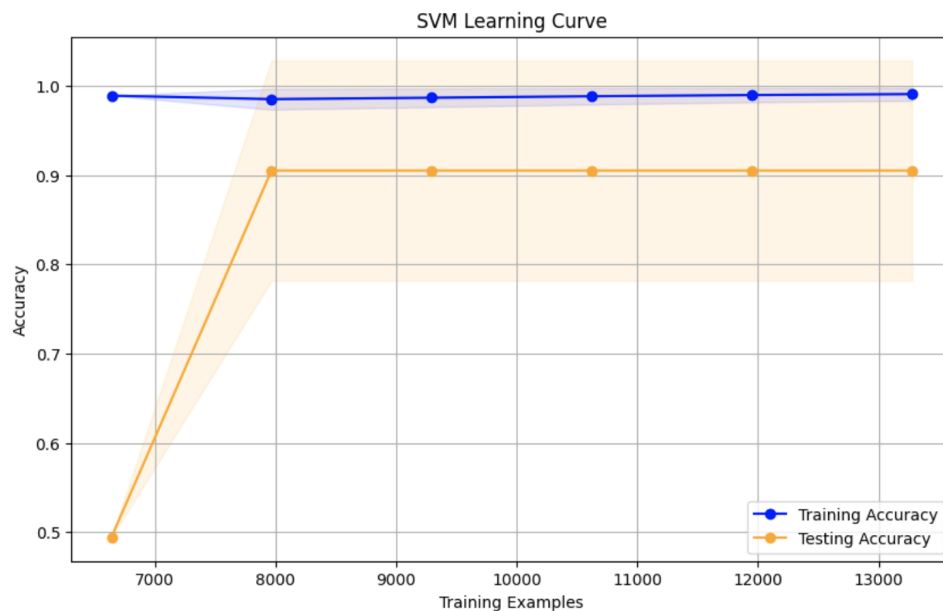


Fig 1: SVM Learning Curve

The KNN performed even better where it scored an accuracy of 1.00 on training

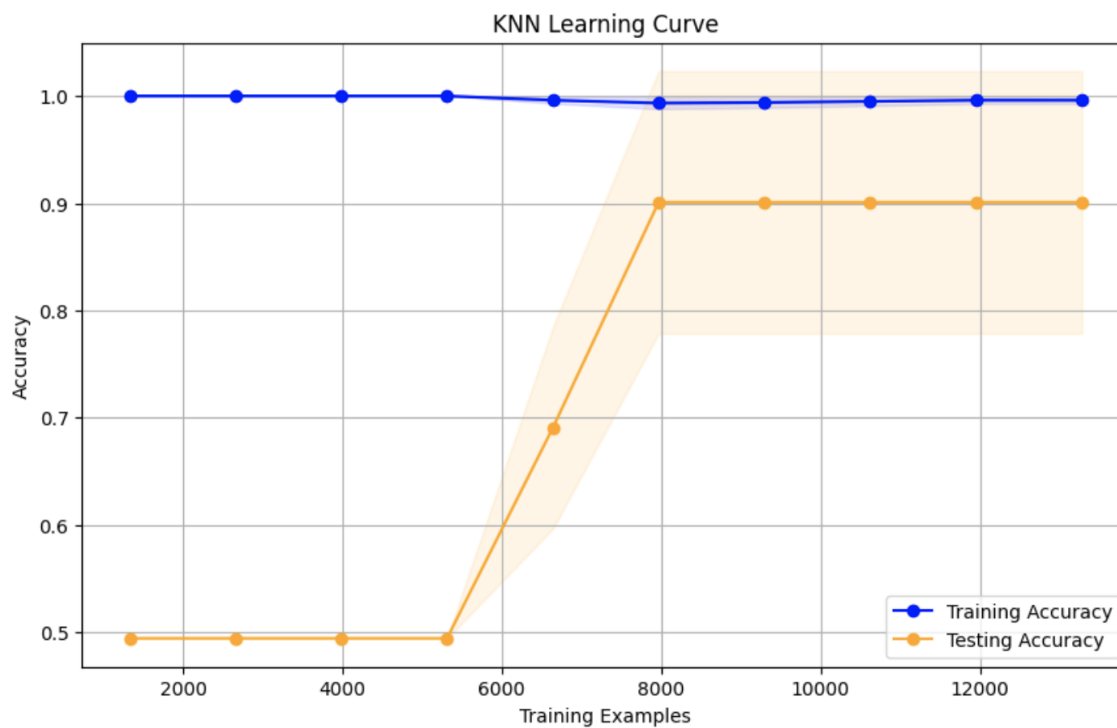


Fig 2: KNN Learning Curve

and 0.99 on the testing dataset. The accuracy started to improve with more than 5000 data points and it improved substantially when trained on over 8000 data points.

Adaboosting algorithm was used which achieved an accuracy score of 0.99.

In case of decision tree, the maximum depth=5 was used.

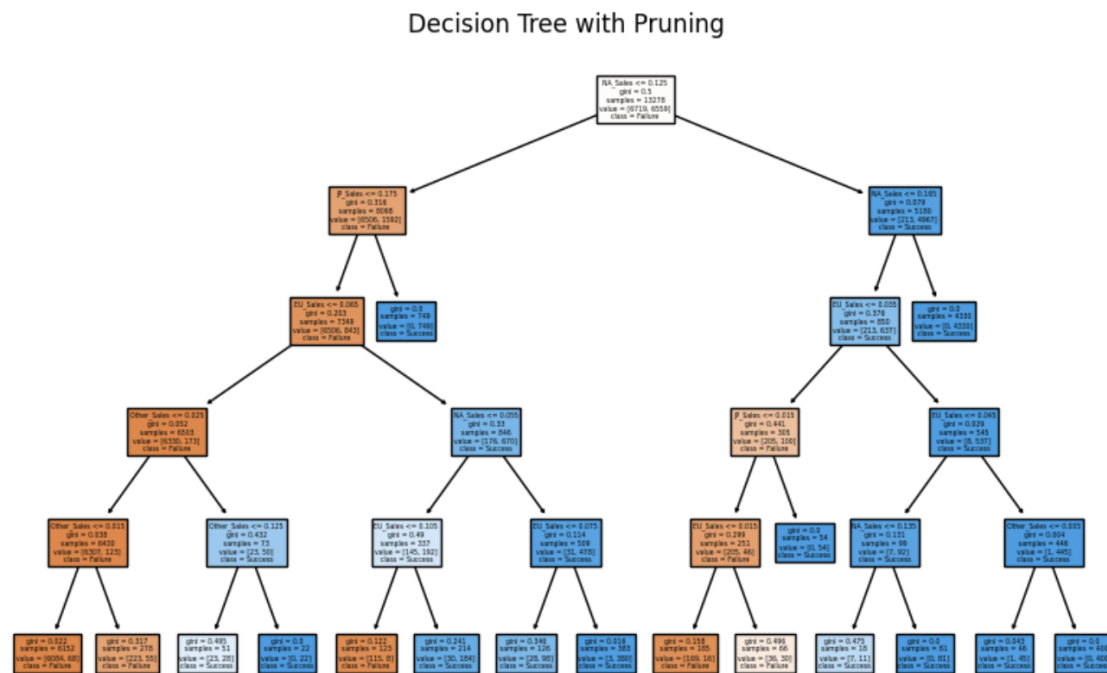


Fig 3: Decision Tree

It achieved an accuracy of 0.98 on both training and testing dataset.

In case of neural network, the 'adam' optimization algorithm was used. The loss function was set to 'binary_crossentropy,' since it was a binary classification problem. The model was trained for 10 epochs and achieved an accuracy of 1.00.

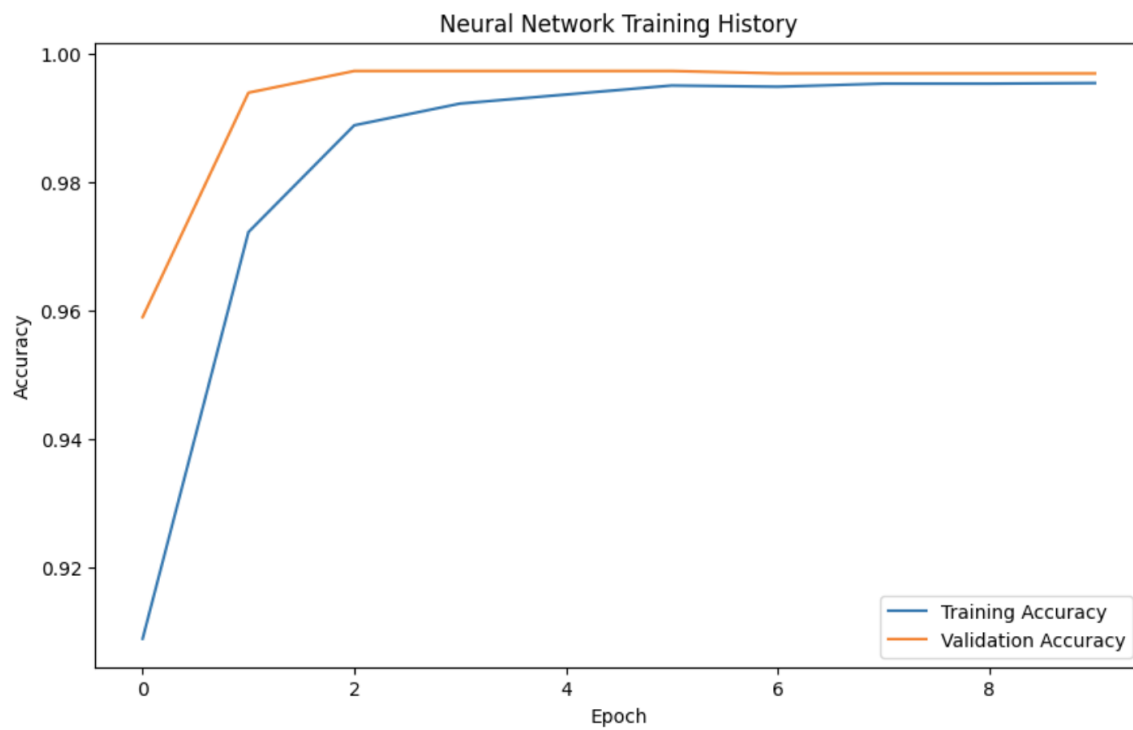


Fig 4: Neural Network Training

Overall, based on the accuracy score the neural network performed the best.