CSCE 5300 – Introduction to Big Data and Data Science PROJECT

Submitted By: Project Group – 3

Towards Award Prediction Based on Big Data Co-Author Network

AUTHORS:

- **Sai Manideep Reddy Pallerla(11553308)**
- **G Tharun Sai Kumar Reddy (11594104)**
- **Varun Chowdary Yarra (11588086)**
- **Prasanth Shabad (11609560)**
- **Pranavi Itharaj(11581907)**

ABSTRACT:

The Academy Awards are presented yearly by the Academy of Motion Picture Arts and Sciences (AMPAS) to celebrate exceptional artistic and technical achievements in the world of cinema. They are regarded globally as a symbol of excellence in the film industry. This dataset contains a full list of nominations and awards in a variety of categories from 1927 through 2020. Although the dataset lacks gender and racial data for nominations and winners, it is a significant resource for investigating the diversity of the Academy Awards. Researchers can examine whether the awards reflect the ethnic variety of films, evaluate the impact of the Academy's new diversity standards, and measure the degree of female representation in the entertainment business. In this study, a model which is sufficient for handling large datasets is implemented for predicting award winners.

INPUT DATASET DESCRIPTION:

We have given a film industry awards dataset from:

Link: https://www.kaggle.com/datasets/dharmikdonga/academy-awards-dataset-oscars

The Oscars, officially known as the Academy Awards, are presented annually to recognize artistic and technical achievements in the film industry. The awards are bestowed by the Academy of Motion Picture Arts and Sciences (AMPAS) and are considered a global recognition of excellence in the field of cinema. To visualize the diversity in the Academy Awards, I searched for a comprehensive dataset containing information on all nominations and winners across categories, along with their respective genders and ethnicities from 1927 to 2020. However, we were only able to find a basic dataset on Kaggle that did not include gender and race data for nominees and winners.

Content Description:

The Dataset contains the Following Data:

- Year_Film: The Year of release of the film.
- Year_Ceremony: The year in which the movie was nominated.
- Ceremony: The Number of Oscars Ceremonies
- Category: The category for which the film was nominated.
- Gender: The gender of the nominee or winner.
- Name: The name of the nominee or winner.
- Race: The racial profile (ethnicity) of the nominee or winner
- Film: The name of the nominated film
- Winner: Indicates whether the entry won (True) or was only nominated (False).
- Acknowledgements: We would like to acknowledge the help of Raphael Fontes, the source of the basic dataset used in this project

Inspiration:

This dataset can be used to explore several questions, such as whether the Academy Awards reflect the racial diversity of films, the reasons behind the introduction of new rules for diversity by the Academy, and the extent of gender representation in the entertainment industry.

INTRODUCTION:

The Academy Awards, also known as the Oscars, are one of the most prestigious awards in the film industry, recognizing excellence in various categories such as Best Picture, Best Director, Best Actor/Actress, and more. With a history that spans nearly a century, the Academy Awards have become a cultural phenomenon, with millions of viewers tuning in each year to watch the ceremony.These are prizes given for artistic and technical excellence in the motion picture industry. The Academy of Motion Picture Arts and Sciences (AMPAS) presents the awards each year as a means of recognizing excellence in cinematic achievements around the world as determined by the voting membership of the Academy and looking for a comprehensive dataset that included all nominees and winners for every category from 1927 through 2020 along with their corresponding gender and ethnicity, but here it was unable to locate one. Based on the method, we gathered a simple dataset from Kaggle, however, the nominees' and winners' racial and gender identities were missing. But, it provides an impactful resource for knowing the diversity of the Academy awards. Finally, this study helped us in exploring the relations in between critical reception, awards recognition and finalizing trends in the film industry by predicting award winners.

DESCRIPTION OF THE MODEL APPROACH:

The problem of analyzing diversity in the Academy Awards can indeed be approached with various decision-making algorithms. Here's a brief overview of the algorithms:

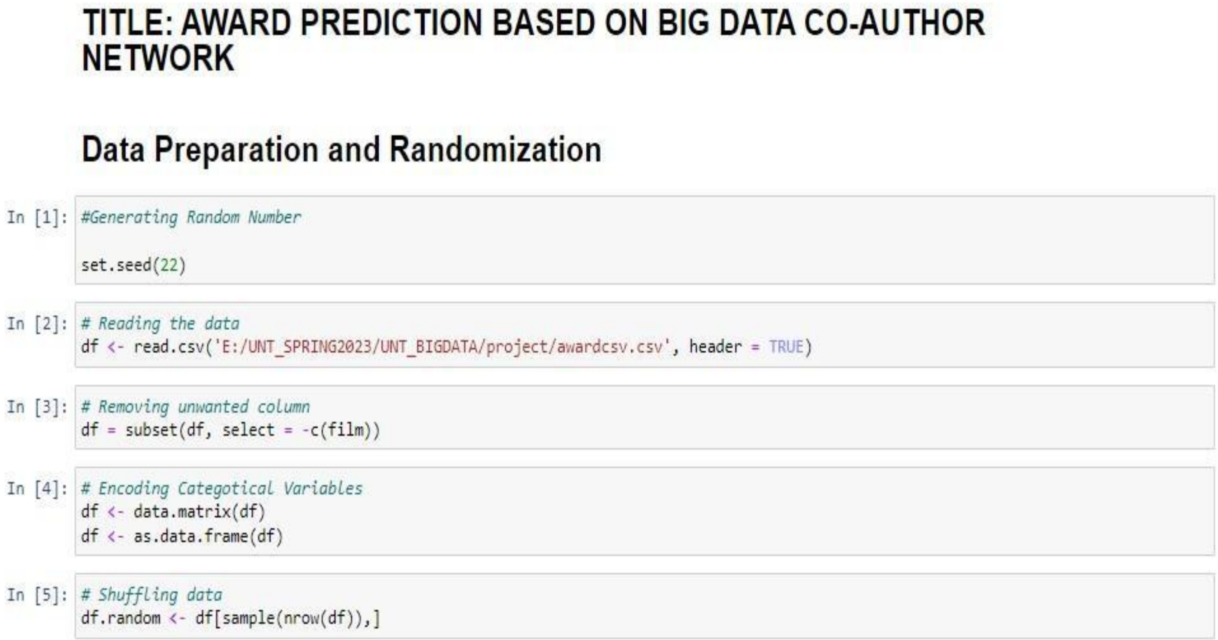
- **Regression:** Regression is a statistical method that models the relationship between a dependent variable and one or more independent variables. In the context of analyzing the diversity in the Academy

Awards, regression could be used to model the relationship between different demographic factors (such as gender, race, or ethnicity) and the likelihood of winning or being nominated for an award..

- **Decision Trees:** Decision trees are a type of algorithm that involves recursively partitioning the input space into smaller subsets based on the values of different features. In the context of analyzing diversity in the Academy Awards, decision trees could be used to identify the most important demographic factors that affect the likelihood of winning or being nominated for an award.
- **Artificial Neural Networks:** Artificial neural networks are a type of machine learning algorithm that are loosely inspired by the structure and function of biological neurons. In the context of analyzing diversity in the Academy Awards, neural networks could be used to model the complex relationships between different demographic factors and the likelihood of winning or being nominated for an award.
- **Feed Forward Neural Networks:** A feedforward neural network is a specific type of artificial neural network where the information flows in only one direction, from input to output. In the context of analyzing diversity in the Academy Awards, a feedforward neural network could be trained to predict the likelihood of winning or being nominated for an award based on demographic factors.

CODE EXECUTION:

- **Step -1:** Data Preparation and Randomization.



TITLE: AWARD PREDICTION BASED ON BIG DATA CO-AUTHOR NETWORK

Data Preparation and Randomization

```
In [1]: #Generating Random Number

        set.seed(22)
```

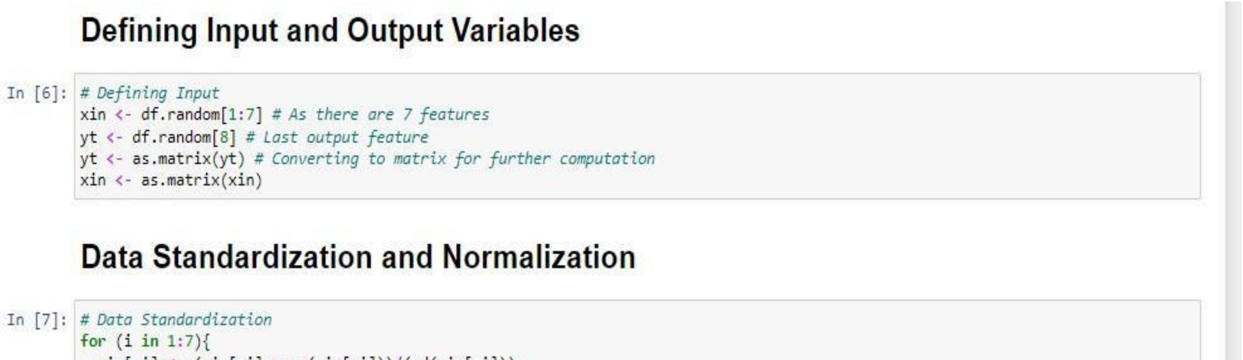
```
In [2]: # Reading the data
        df <- read.csv('E:/UNT_SPRING2023/UNT_BIGDATA/project/awardcsv.csv', header = TRUE)
```

```
In [3]: # Removing unwanted column
        df = subset(df, select = -c(film))
```

```
In [4]: # Encoding Categotical Variables
        df <- data.matrix(df)
        df <- as.data.frame(df)
```

```
In [5]: # Shuffling data
        df.random <- df[sample(nrow(df)),]
```

- **Step -2:** Defining IO Variables and Data Standardization, Normalization**.**



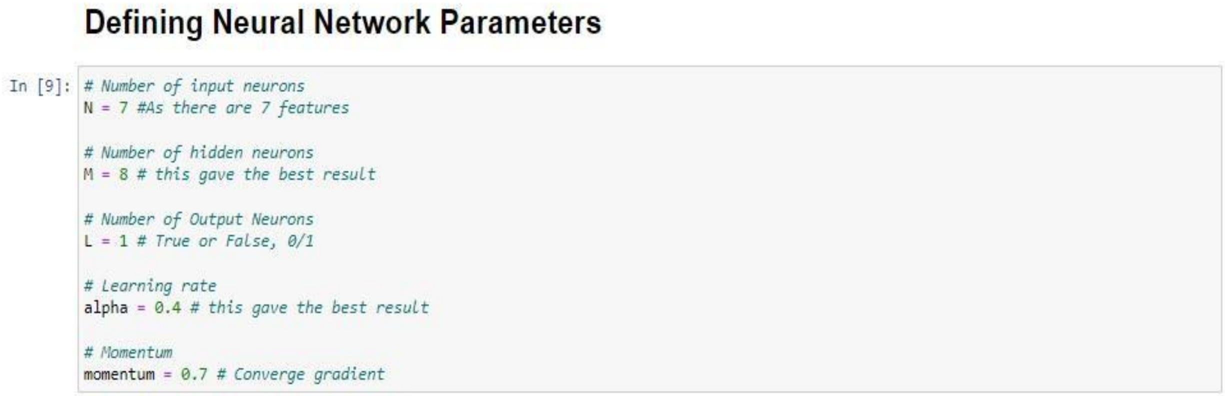
Defining Input and Output Variables

```
In [6]: # Defining Input
        xin <- df.random[1:7] # As there are 7 features
        yt <- df.random[8] # Last output feature
        yt <- as.matrix(yt) # Converting to matrix for further computation
        xin <- as.matrix(xin)
```

Data Standardization and Normalization

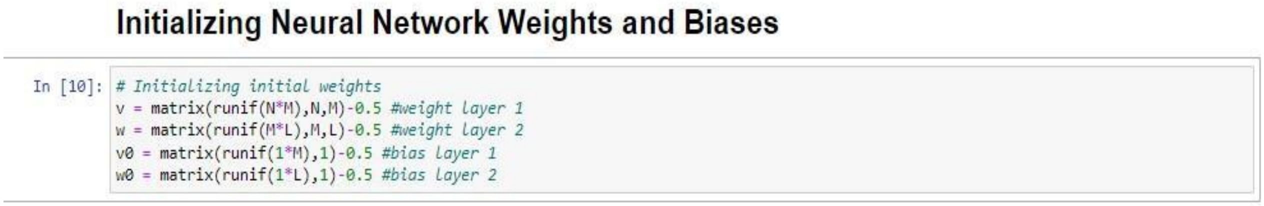
```
In [7]: # Data Standardization
        for (i in 1:7){
```

- **Step -3:** Defining the Neural Network Parameters and Initialization of Neural Network Weights and Biases.



Defining Neural Network Parameters

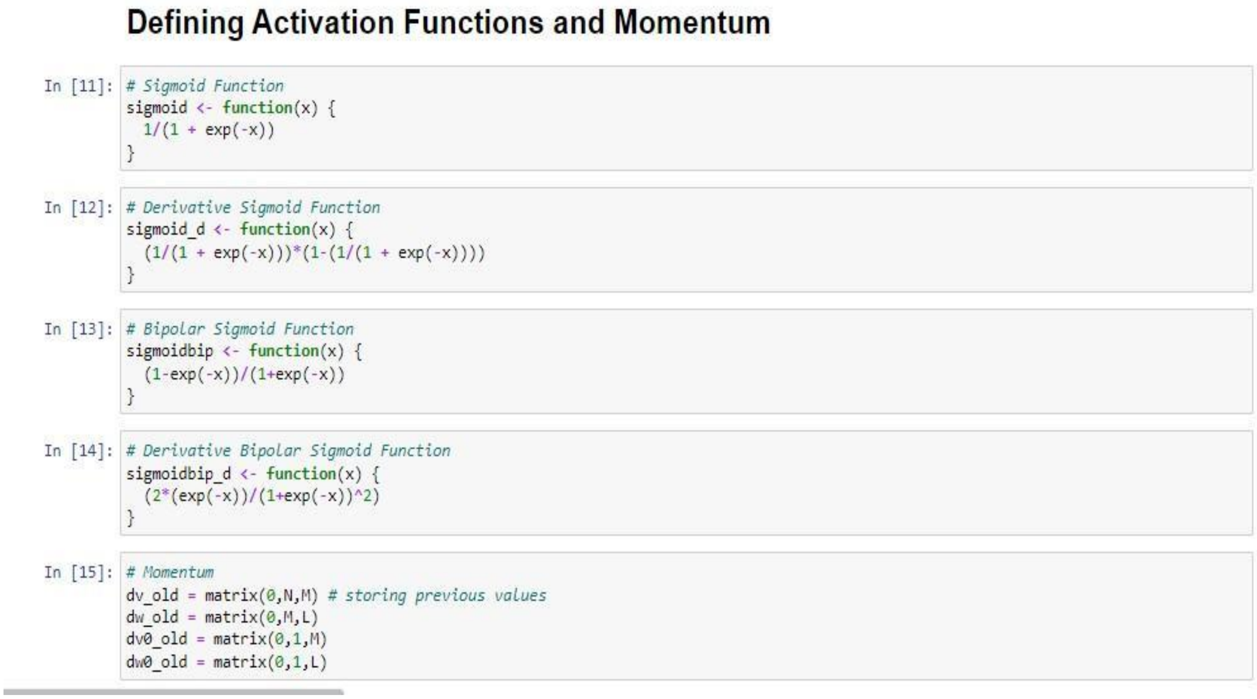
```
In [9]: # Number of input neurons
        N = 7 #As there are 7 features

        # Number of hidden neurons
        M = 8 # this gave the best result

        # Number of Output Neurons
        L = 1 # True or False, 0/1

        # Learning rate
        alpha = 0.4 # this gave the best result

        # Momentum
        momentum = 0.7 # Converge gradient
```



Initializing Neural Network Weights and Biases

```
In [10]: # Initializing initial weights
         v = matrix(runif(N*M),N,M)-0.5 #weight layer 1
         w = matrix(runif(M*L),M,L)-0.5 #weight layer 2
         v0 = matrix(runif(1*M),1)-0.5 #bias layer 1
         w0 = matrix(runif(1*L),1)-0.5 #bias layer 2
```

- **Step -4:** Defining Activation Functions and Momentum**.**



Defining Activation Functions and Momentum

```
In [11]: # Sigmoid Function
         sigmoid <- function(x) {
           1/(1 + exp(-x))
         }
```

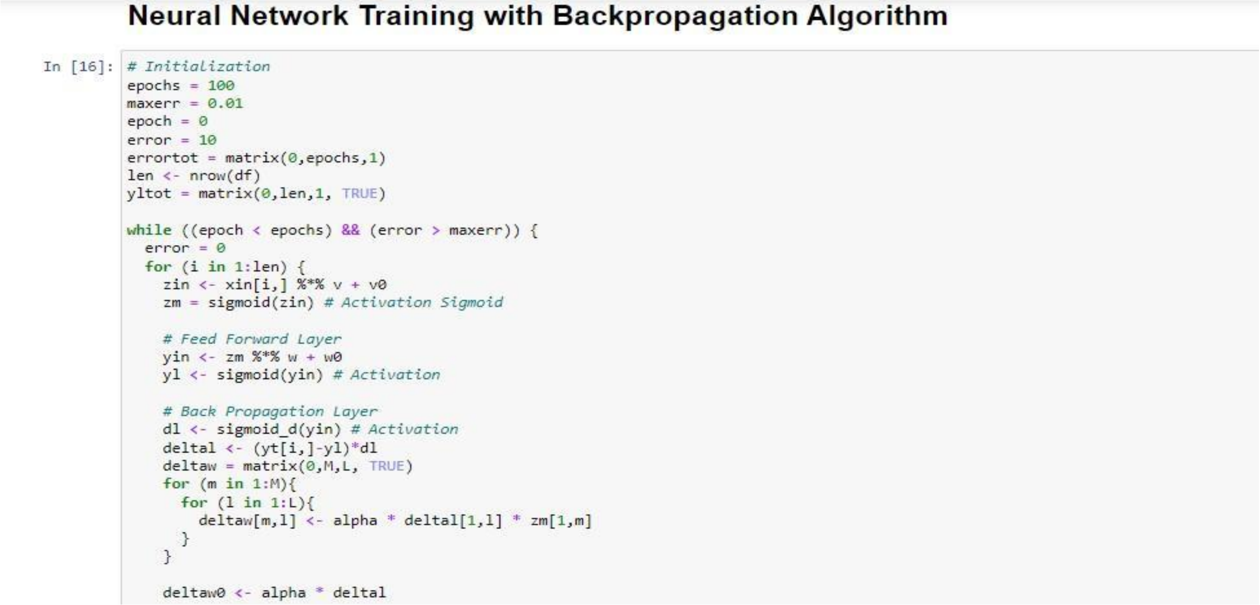
```
In [12]: # Derivative Sigmoid Function
         sigmoid_d <- function(x) {
           (1/(1 + exp(-x)))*(1-(1/(1 + exp(-x))))
         }
```

```
In [13]: # Bipolar Sigmoid Function
         sigmoidbip <- function(x) {
           (1-exp(-x))/(1+exp(-x))
         }
```

```
In [14]: # Derivative Bipolar Sigmoid Function
         sigmoidbip_d <- function(x) {
           (2*(exp(-x))/(1+exp(-x))^2)
         }
```

```
In [15]: # Momentum
         dv_old = matrix(0,N,M) # storing previous values
         dw_old = matrix(0,M,L)
         dv0_old = matrix(0,1,M)
         dw0_old = matrix(0,1,L)
```

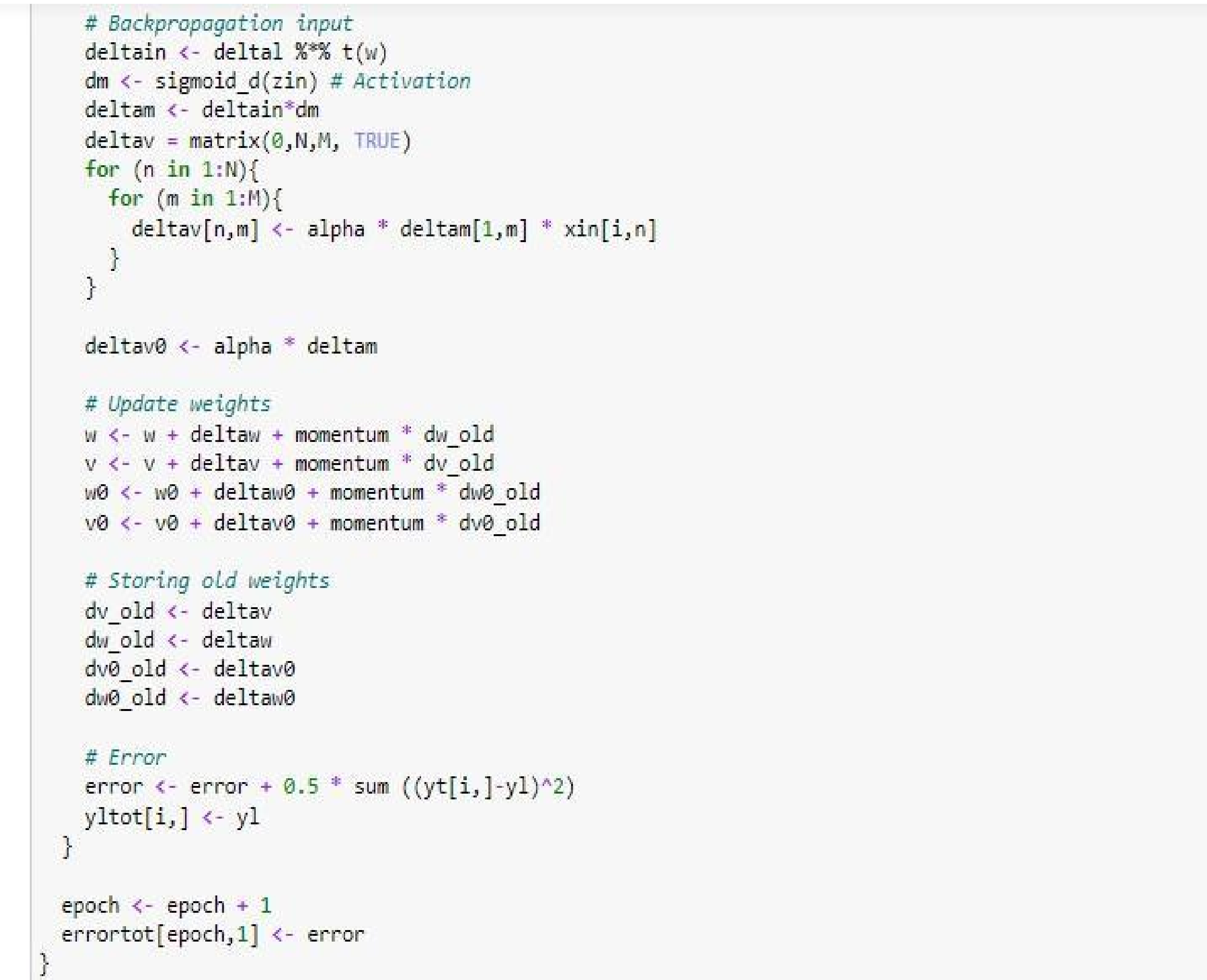
- **Step -5:** Neural Network Training with Backpropagation Algorithm**.**



Neural Network Training with Backpropagation Algorithm

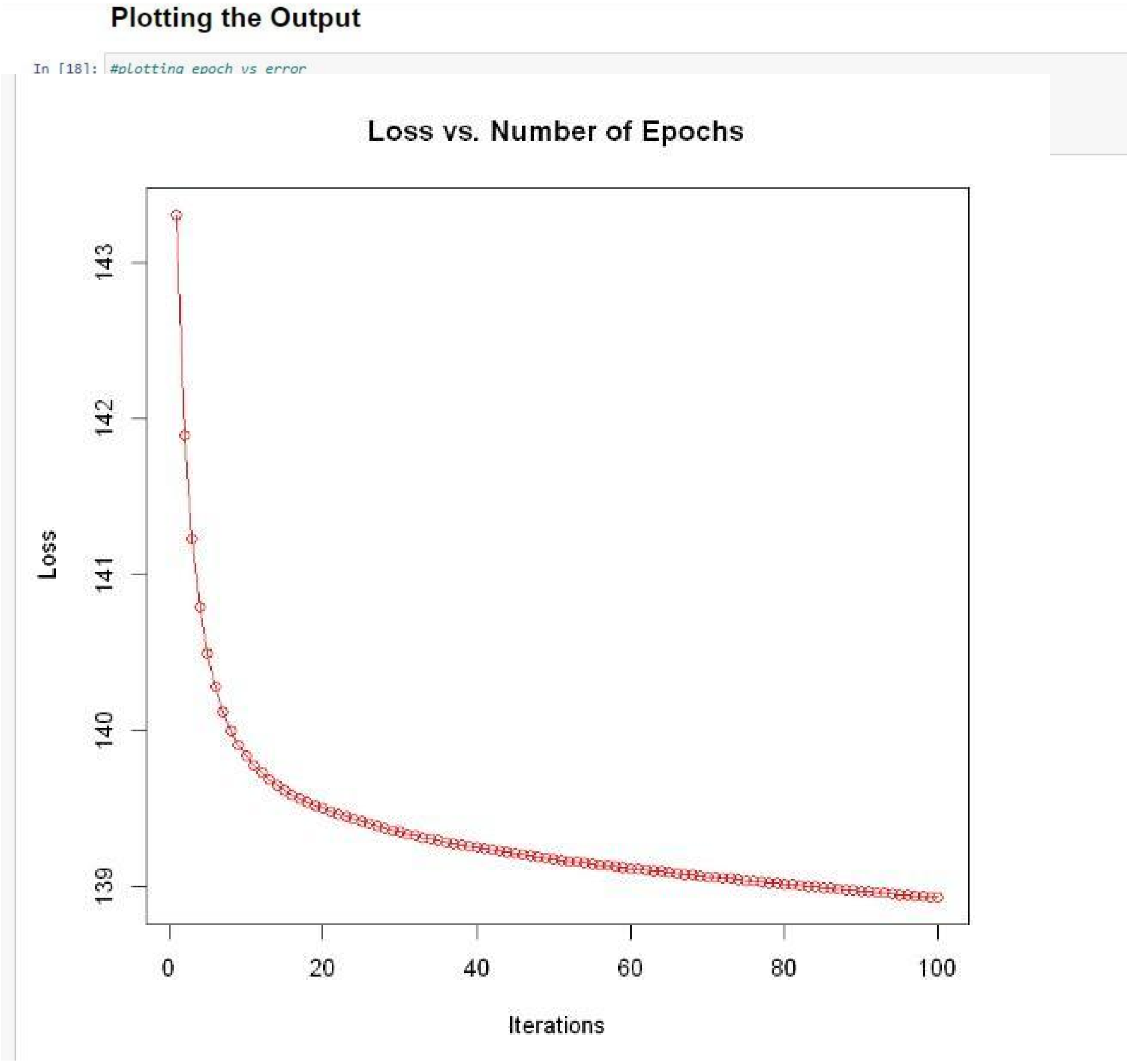
In [16]:

```
# Initialization
epochs = 100
maxerr = 0.01
epoch = 0
error = 10
errortot = matrix(0,epochs,1)
len <- nrow(df)
yltot = matrix(0,len,1, TRUE)

while ((epoch < epochs) && (error > maxerr)) {
  error = 0
  for (i in 1:len) {
    zin <- xin[i,] %*% v + v0
    zm = sigmoid(zin) # Activation Sigmoid

    # Feed Forward Layer
    yin <- zm %*% w + w0
    yl <- sigmoid(yin) # Activation

    # Back Propagation Layer
    dl <- sigmoid_d(yin) # Activation
    deltal <- (yt[i,]-yl)*dl
    deltaw = matrix(0,M,L, TRUE)
    for (m in 1:M){
      for (l in 1:L){
        deltaw[m,l] <- alpha * deltal[1,l] * zm[1,m]
      }
    }

    deltaw0 <- alpha * deltal
```



```
    # Backpropagation input
    deltain <- deltal %*% t(w)
    dm <- sigmoid_d(zin) # Activation
    deltam <- deltain*dm
    deltav = matrix(0,N,M, TRUE)
    for (n in 1:N){
      for (m in 1:M){
        deltav[n,m] <- alpha * deltam[1,m] * xin[i,n]
      }
    }

    deltav0 <- alpha * deltam

    # Update weights
    w <- w + deltaw + momentum * dw_old
    v <- v + deltav + momentum * dv_old
    w0 <- w0 + deltaw0 + momentum * dw0_old
    v0 <- v0 + deltav0 + momentum * dv0_old

    # Storing old weights
    dv_old <- deltav
    dw_old <- deltaw
    dv0_old <- deltav0
    dw0_old <- deltaw0

    # Error
    error <- error + 0.5 * sum ((yt[i,]-yl)^2)
    yltot[i,] <- yl
  }

  epoch <- epoch + 1
  errortot[epoch,1] <- error
}
```

- **Step -6:** Plotting the Output.



Plotting the Output

```
In [18]: #plotting epoch vs error
```

OBTAINED RESULTS:

Results show the Number of iterations performed and the loss function, In comparison with the IEEE paper considered for the implementation we can see that the Y-axis being for loss function has more units and the curve changes more drastically for increase in the number of iterations. It is easily observed that the loss function of the prediction model using the BPNN algorithm causes the loss function to decrease.

INDIVIDUAL CONTRIBUTION:

These are the individual contributions done by the team members by referring to the IEEE journals and surfing through the web and gathered the required information .

NAME	**CONTRIBUTION**
Sai Manideep Reddy Pallerla	Code Execution, Obtained Results
Gonnuru Tharun sai kumar reddy	Introduction, Advantages and Limitations of the Approach
Varun Chowdary Yarra	Practical Applications of the Approach, Conclusion
Prasanth Shabad	Abstract ,References
Pranavi Itharaj	Input Data , Decsription of the model approach

ADVANTAGES AND LIMITATIONS OF THE APPROACH:

Advantages:

- **Can work with any number of outputs**: The analysis of the Academy Awards dataset can involve examining different aspects of diversity, such as gender, race, and ethnicity. The analysis can look at any number of output variables, depending on the research question being asked when compared with the considered Towards award prediction.
- **Feature importance analysis:** Analyzing the Academy Awards dataset can help identify which demographic factors are most important in determining who wins or is nominated for awards. For example, the analysis may reveal whether gender or race has a greater impact on the likelihood of winning an award depending on the threshold set and .
- **Simple, fast, and easy to implement for a larger data:** Depending on the complexity of the analysis, analyzing the Entertainment Awards dataset can be relatively simple and straightforward. The dataset is well-structured and contains clearly defined variables, making it easy to analyze using statistical software. Additionally, the dataset covers a large timespan, allowing for longitudinal analyses to be

conducted, which when compared with the considered paper is typically not so easier to implement due to their consideration being.

Limitations:

- **Performance is dependent on input data:** This limitation applies to many machine learning algorithms, not just decision trees. The accuracy and effectiveness of any algorithm can be highly dependent on the quality of the input data. If the input data is incomplete, inconsistent, or biased, the algorithm may not be able to accurately represent the relationships between the input variables and the output.
- **Sensitive to noisy data:** Many machine learning algorithms, including decision trees, can be sensitive to noisy data or outliers in the input data. Outliers or data points that are significantly different from the rest of the data can affect the accuracy of the model and lead to overfitting.
- **Does not support batch approach:** This limitation may apply to some machine learning algorithms, particularly those that require a sequential or iterative approach to model building. Batch processing, which involves processing large amounts of data in a parallelized or distributed manner, may not be well-suited for some algorithms, and may require specialized approaches or hardware to achieve efficient processing.

PRACTICAL APPLICATIONS OF THE APPROACH:

- **Language Processing:** Language processing is a branch of artificial intelligence that deals with the interactions between computers and human languages. This field encompasses both natural language processing (NLP) and natural language generation (NLG). NLP involves analyzing and understanding human language, while NLG involves generating natural language by machines. Language processing has numerous applications, such as chatbots, virtual assistants, search engines, sentiment analysis, and language translation. The techniques used in language processing include statistical models, machine learning, rule-based systems, and deep learning. However, one of the main challenges in language processing is dealing with the complexity and ambiguity of human language, which can vary widely depending on context, dialect, and other factors.
- **Image Recognition:** Image recognition refers to the ability of computers to identify and classify objects in digital images. This involves utilizing computer algorithms and machine learning techniques to analyze patterns and features within images. Image recognition has many applications, such as self-driving cars, facial recognition, medical imaging, security, and e-commerce. Image recognition algorithms can be trained using supervised learning, unsupervised learning, or reinforcement learning. However, one of the main challenges in image recognition is dealing with variations in lighting, perspective, and background, which can affect the accuracy of the algorithm. Additionally, large amounts of labeled data are required to train the algorithm effectively.
- **Medical Diagnostic**: Medical diagnostic involves using medical data to diagnose diseases or conditions in patients. This involves analyzing patient symptoms, medical history, and test results to make a

diagnosis. Medical diagnostic techniques include machine learning, statistical models, and rule-based systems. Medical diagnostic has numerous applications, such as radiology, pathology, cardiology, and genomics. However, one of the main challenges in medical diagnostic is dealing with the complexity and variability of medical data, which can be affected by factors such as patient age, gender, and lifestyle. Another challenge is ensuring the accuracy and reliability of the diagnosis, which can be affected by the quality of the data and the expertise of the medical professionals involved.

- **Time Series Prediction:** Time series prediction is the process of forecasting future values in a time series based on past observations. This involves using statistical models, machine learning algorithms, and other techniques to analyze patterns and trends within the data. Time series prediction has numerous applications, such as financial forecasting, weather forecasting, and stock market analysis. Time series prediction techniques include autoregression, moving averages, exponential smoothing, and neural networks. However, one of the main challenges in time series prediction is dealing with the complexity and variability of time series data, which can be affected by factors such as seasonality, trends, and noise. Another challenge is ensuring the accuracy and reliability of the prediction, which can be affected by the quality of the data and the complexity of the model.
- **Signature Verification:** Signature verification is the process of verifying the authenticity of a signature, often used in security and fraud prevention. This involves analyzing signatures and identifying patterns to determine whether a signature is genuine or forged. Signature verification techniques include machine learning, image processing, and rule-based systems. Signature verification has numerous applications, such as banking, legal, and government, where signatures are used for identity verification and document authentication. However, one of the main challenges in signature verification is dealing with variations in signatures, which can be affected by factors such as age, health, and emotional state. Another challenge is ensuring the accuracy and reliability of the verification process, which can be affected by the quality of the data and the expertise of the analysts involved.

CONCLUSION:

In conclusion, this study concludes that the Academy Awards dataset which we are taking provides a rich and comprehensive resource for studying the history and trends of the Academy Awards. With information on all nominees and winners from the first ceremony in 1927 to the most recent ceremony in 2020, the dataset offers a valuable tool for researchers interested in exploring the relationship between critical reception and awards recognition, predicting award winners, and analyzing trends in the film industry over time. Whether you are a film buff, a data analyst, or a researcher, this dataset is a great starting point for diving deeper into the world of the Academy Awards and understanding how they have evolved over the years. In both cases, the loss is decreasing as the number of epochs increases.

REFERENCES:

- **https://ieeexplore.ieee.org/document/8725612**
- **https://www.medsci.cn/sci/show_paper.asp?id=29a00121ce5071de**
- **https://epjdatascience.springeropen.com/articles/10.1140/epjds/s13688-014-0009-x**
- **https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00432-y**
- Yan, E., & Ding, Y. (2009). Applying Centrality Measures to Impact Analysis: A Coauthorship Network Analysis. Journal of the American Society for Information Science and Technology, 60(10), 2107-2118.
- Yang, Z., Yin, D., & Davison, B. D. (2011, July). Award Prediction with Temporal Citation Network Analysis. In Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 1203-1204). ACM.
-] Chuan, T., Juan, T., Junmin, F., et al. (2015). Study on the Identifiability and Predictability of Turing Award Winners Based on Multiple Bibliometric Indicators and Support Vector Machine. Journal of Intelligence.
- Newman, M. E. (2001). The Structure of Scientific Collaboration Networks. Proceedings of the National Academy of Sciences, 98(2), 404-409 Newman, M. E. (2001). The Structure of Scientific Collaboration Networks. Proceedings of the National Academy of Sciences, 98(2), 404-409.
- **https://datalab.ucdavis.edu/2019/08/27/creating-co-author-networks-in-r/ -**