Performance Comparison of Neural Networks and Traditional Regression Models on Diverse Datasets

David Bishop DBISHOP7@JH.EDU
Computer Scientist
John's Hopkins University
Washington, DC 20001, USA

Editor: David Bishop

Abstract

This study evaluates the performance of Combined Neural Networks, Feed Forward Networks, and Linear/Logistic Regression models across six diverse datasets: Abalone, Breast Cancer, Car Evaluation, Congressional Voting Records, Computer Hardware, and Forest Fires. By examining classification and regression tasks, this research aims to identify the most effective model type for each dataset. The results indicate that neural network-based models, particularly the Combined and Feed Forward Networks, consistently outperform traditional Linear/Logistic Regression models, especially in complex and multi-class scenarios. Key observations highlight the variability in model performance based on dataset characteristics, emphasizing the importance of model selection tailored to the nature of the data. This paper underscores the strengths of neural networks in modern machine learning tasks and suggests avenues for future research, including refining neural network architectures and exploring advanced preprocessing techniques to further enhance model performance.

Keywords: Nerual Networks, Linear Regression, Logistic Regression, Combined Nerual Networks, Feed Forward Networks, Classification, Regression

1 Introduction

Machine learning encompasses a wide array of tasks, with classification and regression being among the most fundamental (7). These tasks are crucial in diverse applications, ranging from medical diagnosis to financial forecasting. In this project, we explore the performance of various machine learning models—specifically Combined Neural Networks, Feed Forward Networks, and Linear/Logistic Regression models—across multiple datasets to evaluate their efficacy and determine the best-performing model type.

Neural networks have become increasingly prominent in machine learning due to their ability to model complex patterns and relationships within data (8). Combined Neural Networks, which integrate multiple network structures, and Feed Forward Networks, a type of deep neural network, offer significant advantages in capturing these intricate patterns. Conversely, Linear and Logistic Regression models, while simpler and more interpretable, often struggle with more complex datasets but can still provide valuable insights due to their straightforward nature.

The primary objective of this study is to compare the performance of these models on six distinct datasets: Abalone, Breast Cancer, Car Evaluation, Congressional Voting Records

(House Votes 84), Computer Hardware (Machine), and Forest Fires. Each dataset presents unique challenges, ranging from multi-class classification to complex regression tasks, thus providing a comprehensive evaluation of each model's strengths and limitations.

Our hypothesis states that the Combined Neural Network will outperform the other models due to its integrated approach, followed closely by the Feed Forward Network. We anticipate that Linear/Logistic Regression models will lag behind, particularly on more complex datasets, due to their simplicity.

To test this hypothesis, we implemented each model and evaluated their performance using relevant metrics for both regression (Mean Squared Error and R-squared) and classification (Accuracy, Precision, Recall, and F1 Score). The experiments were conducted using robust cross-validation techniques to ensure the reliability and generalizability of the results.

This paper is structured as follows: Section 2 details the algorithms and experimental methods, including preprocessing steps and dataset descriptions. Section 3 presents the experimental results, highlighting the performance of each model across the different datasets. Section 4 provides a detailed discussion and interpretation of these results, and Section 5 concludes the paper, summarizing the findings and suggesting directions for future research.

2 Algorithms and Experimental Methods

During this project, we utilized various tools and techniques to create, run, and validate: Linear and Logistic Regression Models, Feedforward Neural Networks, and Combined Networks utilizing an autoencoder.

2.1 Experimental Approach

For each dataset, the initial step involved loading and preprocessing the data. This included handling missing values, normalizing features, and converting categorical variables to numerical representations where necessary. These steps are detailed in Section 2.3.

Once the datasets were prepared, they were split into training and testing sets using an 80-20 split. This means 80% of the data was used for training the model, while 20% was reserved for testing its performance. Each model was then trained on the training data, allowing it to learn from the features and target values.

To ensure robust evaluation of model performance, we employed 5x2 cross-validation. This method involves splitting the data into five folds and repeating the process twice. By averaging the results across different subsets of the data, cross-validation helps mitigate the risk of overfitting and provides a more reliable estimate of the model's performance.

After running the cross-validation we calculated a few different performance metrics to ensure proper ranking of the models. These metrics are discussed in section 3.

2.2 datasets

As discussed earlier in the paper we utilized six separate datasets to see the functionality of the Decision Tree for classification and regression

2.2.1 Abalone

The Abalone dataset consists of physical measurements and attributes of abalone, a type of marine mollusk. This dataset is intended for regression problems, as the target variable is the number of rings on the abalone's shell, which can be used to estimate its age. The dataset includes features such as length, diameter, height, whole weight, shucked weight, viscera weight, shell weight, and sex. The challenge is to predict the number of rings based on these physical characteristics.

2.2.2 Breast Cancer

The Breast Cancer dataset describes characteristics of cell nuclei present in images of breast cancer biopsies. This dataset is intended for classification problems, with the target variable being the diagnosis (M for malignant, B for benign). The dataset includes features such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. Each feature is a real-valued number.

2.2.3 Car Evaluation

The Car Evaluation dataset provides evaluations of car acceptability based on price, comfort, and technical specifications. This dataset is used for classification problems, with the target variable being the acceptability of the car (unacc, acc, good, vgood). The dataset includes features such as buying price, maintenance price, number of doors, capacity in terms of persons to carry, the size of luggage boot, and safety.

2.2.4 Congressional Vote

The Congressional Vote dataset includes votes for each of the U.S. House of Representatives Congressmen on 16 key votes. This dataset is used for classification problems, with the target variable being the party affiliation (democrat or republican). The dataset includes features representing votes on key issues, recorded as 'yes', 'no', or 'abstain'.

2.2.5 Computer Hardware

The Computer Hardware dataset describes the relative CPU performance of computer hardware based on features such as cycle time and memory size. This dataset is used for regression problems, with the target variable being the performance (PRP). The dataset includes features such as machine cycle time, memory size, cache size, and channel count.

2.2.6 Forest Fires

The Forest Fires dataset involves predicting the burned area of forest fires using meteorological and other data. This dataset is used for regression problems, with the target variable being the burned area. The dataset includes features such as the x and y spatial coordinates, month, day, FFMC, DMC, DC, ISI indices from the FWI system, temperature, relative humidity, wind, and rain.

2.3 Data Pre-processing

During this project there were a couple steps of pre-processing that needed to happen. For the Breast Cancer dataset we needed to map the benign and malignant (2 and 4) into a one or zero to ensure the models could accurately predict the values. For any dataset that utilized letters or words instead of numbers we needed to encode into int types to ensure the model could process the data correctly and efficiently. One of the more important pre-processing steps to mention was done to the Forest Fires dataset. Due to the large amount of zero values contained in the dataset we decided to utilize log so the model did not always predict zero.

3 Results

This section will be displaying the results of each models performance. The performance of the models were assessed using various metrics tailored to the type of task. For regression tasks, we utilized Mean Squared Error (MSE) and R-squared (R^2) scores. MSE measures the average squared difference between the observed actual outcomes and the outcomes predicted by the model, providing insight into the average magnitude of prediction errors, with lower values indicating better performance. R^2 score provides the proportion of variance in the dependent variable predictable from the independent variables, with values closer to 1 indicating a better fit.

For classification tasks, we used Accuracy, Precision, Recall, and F1 scores. Accuracy measures the proportion of correctly predicted instances out of the total instances, providing a straightforward indication of overall model performance. Precision, the ratio of true positive predictions to the total positive predictions, and Recall, the ratio of true positive predictions to the actual positives, are crucial for evaluating models in imbalanced datasets. The F1 score, the harmonic mean of Precision and Recall, offers a balanced measure that considers both false positives and false negatives, making it ideal for cases with uneven class distributions. Below are these metrics compared between the four models, attached at the bottom will be the generated Confusion Matrix's for Classification datasets and Prediction Graphs for Regression Datasets.

3.1 Abalone

Metric	Linear Regression	Combined Neural Network	Feed Forward Network
Mean Squared Error	5.4660	0.0076	4.8112
R^2 Score	0.4737	0.4419	0.5368

3.2 Breast Cancer

Metric	Logistic Regression	Combined Neural Network	Feed Forward Network
Accuracy	0.9606	0.9628	0.9866
Precision	0.9336	0.9624	0.9866
Recall	0.9515	0.9632	0.9866
F1 Score	0.9420	0.9628	0.9866

3.3 Car Evaluation

Metric	Logistic Regression	Combined Neural Network	Feed Forward Network
Accuracy	0.0398	0.9469	0.9783
Precision	0.1513	0.9049	0.9566
Recall	1.0000	0.8795	0.9566
F1 Score	0.2605	0.8919	0.9566

3.4 Computer Hardware

Metric	Linear Regression	Combined Neural Network	Feed Forward Network
Mean Squared Error	0.0032	0.0021	0.0016
R^2 Score	0.8048	0.8873	0.8865

3.5 Congressional Voting Records

Metric	Logistic Regression	Combined Neural Network	Feed Forward Network
Accuracy	0.8831	0.9438	0.9798
Precision	0.8571	0.9438	0.9798
Recall	0.8433	0.9438	0.9798
F1 Score	0.8454	0.9438	0.9798

3.6 Forest Fires

Metric	Linear Regression	Combined Neural Network	Feed Forward Network
Mean Squared Error	0.0392	0.0398	0.0414
R^2 Score	-0.0355	0.0144	-0.0486

4 Discussion

The previous section presented the performance of the models across six distinct datasets. In this section, we will look deeper and analyze these results, interpreting the data points and exploring the implications of our findings. Each subsection will discuss the performance on a specific dataset, highlighting key observations and potential areas for further improvement.

4.1 Abalone

Based on the scores the combined neural network managed to perform the best in both MSE and R^2. Both the Linear Regression model and the Feed Forward Network failed to perform according to the MSE and performed very poorly with their R^2 scores. Looking at generated graphs (figures 1, 2, 3) the predictions made by the linear regressor and feed forward are much more spread out. However, they still do follow the perfect predictions line reasonably well. Removing some of the unnecessary columns or possibly formatting the data differently may have helped these models to perform better.

4.2 Breast Cancer

Each of the models performed exceedingly well on the Breast Cancer dataset. The Feed Forward Network achieved the best scores which is furthered by (Figure 5) showing that it performed perfectly (on the dataset that it was run to generate the confusion matrix). The Combined Neural Network performed slightly lower but still performing incredibly well(Figure 4). The Logistic Regression model performed the worst overallbased on its performance metrics but did seem to perform well on the confusion matrix (Figure 6). This high performance indicates that the models were well-suited for the classification task, most likely due to the relatively simple nature of the dataset and the clear distinction between classes.

4.3 Car Evaluation

For the Car Evaluation dataset, the Feed Forward Network outperformed the other models (Figure 8). The Combined Network also performed well (Figure 7). Both models missed only a percentage of two of predictions according to their confusion matricies. However, the Logistic Regression model performed poorly in every aspect except recall, this makes sense when looking at the confusion matrix (Figure 9) displaying that it predicted a true label 77 times where it should have been false. If the other models had performed poorly we would attribute this to a bad dataset, but as the other models performed fairly well this is most likely due to the multi-class nature of the problem, which tends to be better handled by more complex models like neural networks.

4.4 Computer Hardware

For the Machine dataset, the Combined Network achieved the best performance of the models, its predicted vs actual chart (Figure 16) truly displays this as the dots fall nearly perfectly on the line. The Feed Forward Network performed slightly worse but the graph (Figure 17) still shows that it was capable of understanding the pattern. The Linear Regression model performed alright but as displayed on the graph (Figure 18) was a little more sporadic in its predictions. The lower performance of the linear regression model may be due to it not fully understanding the feature interactions, something that neural nets are much more adept at doing.

4.5 Congressional Voting Records

For the House Votes 84 dataset, the Feed Forward Network performed the best with its Confusion Matrix (Figure 14) getting a perfect score. The Combined Network also performed very strongly with only 9 incorrect predictions (Figure 13). The Logistic Regression model performed fairly well but statistically slightly worse than the combined network (Figure 15). These results suggest that neural networks, especially the Feed Forward Network, are better suited for this classification task compared to logistic regression.

4.6 Forest Fires

The Forest Fires dataset posed a challenge for all models, but the Feed Forward Network achieved the best performance. Looking at its graph (Figure 11) we can see that unlike the other two models it seemed to understand the relationship of the data a little better. The Combined Network also understood the data but not quite as well (Figure 10). The Linear Regression model performed the worst and its predicted datapoints (Figure 12) show that as they seem the most sporadic. The relatively high MSE and low R^2 scores indicate that the models struggled to make accurate predictions, likely due to the complexity and variability of the dataset. As mentioned in an earlier section the forest fires dataset is heavily weighted towards zero. This can be seen with the drastic amount of low predictions made by the models in the referenced figures.

5 Conclusion

The findings of this study align with our initial hypothesis that neural network-based models, particularly the Combined and Feed Forward Networks, would outperform traditional Linear/Logistic Regression models. The performance advantage of neural networks is evident in most datasets, especially those with more complex structures and multi-class classifications.

One key observation is the variability in model performance based on dataset characteristics. For simpler datasets like Breast Cancer, even traditional regression models can achieve high accuracy, whereas more complex datasets like Forest Fires necessitate advanced neural network models to capture intricate patterns effectively.

The results underscore the importance of selecting appropriate models based on the dataset's nature and complexity.Future work could focus on refining neural network architectures, exploring advanced preprocessing techniques, and applying these models to a broader range of datasets. Additionally, investigating ensemble methods that combine the strengths of different models could provide further performance improvements.

Overall, this investigation highlights the strengths of neural networks in modern machine learning tasks, suggesting that continued advancements in neural network methodologies will likely lead to even greater performance gains across various applications.

References

[1] Marine Research Laboratories, Tasmania, 1995. Abalone Dataset https://archive.ics.uci.edu/ml/datasets/Abalone

[2] Wolberg,WIlliam. (1992). Breast Cancer Wisconsin (Original). UCI Machine Learning Repository. https://doi.org/10.24432/C5HP4Z

[3] Jozef Stefan Institute, Yugoslavia (Slovenia), 1988. Car Evaluation Dataset https://archive.ics.uci.edu/ml/datasets/Car+Evaluation

[4] University of California, Irvine, 1987, Congressional Voting Records Dataset https://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records

[5] Tel Aviv University, Israel, 1987. Computer Hardware Dataset https://archive.ics.uci.edu/ml/datasets/Computer+Hardware

[6] University of Minho, Portugal, 2007. Forest Fires Dataset https://archive.ics.uci.edu/ml/datasets/Forest+Fires

[7] Bobbitt, Z. (2021, June 22). Regression vs. Classification: What's the Difference? Statology. https://www.statology.org/regression-vs-classification/

[8] Cuomo, S., Di Cola, V. S., Giampaolo, F., Rozza, G., Raissi, M., & Piccialli, F. (2022). Scientific Machine Learning Through Physics–Informed Neural Networks: Where we are and What's Next. Journal of Scientific Computing, 92(3). https://doi.org/10.1007/s10915-022-01939-z

6 Referenced Graphics

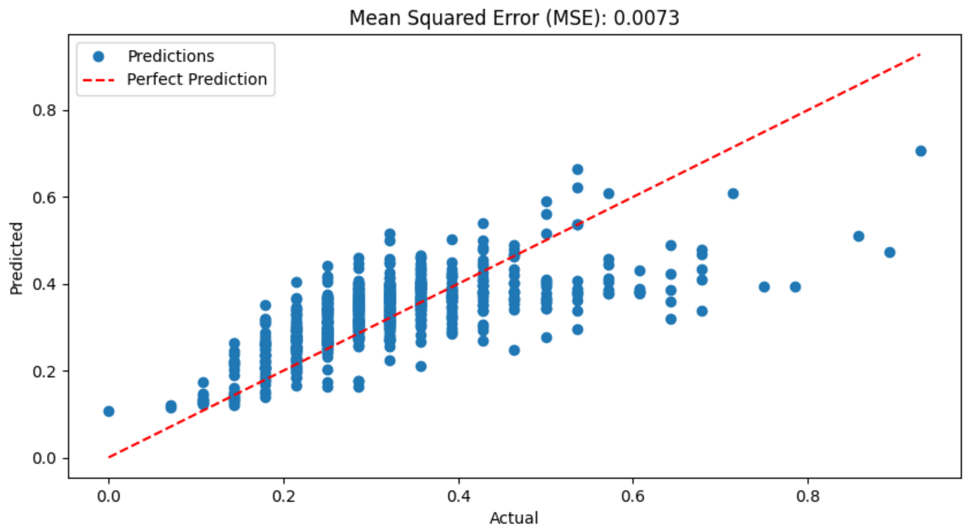

Figure 1: Mean Squared Error plot for Abalone dataset - Combined Network

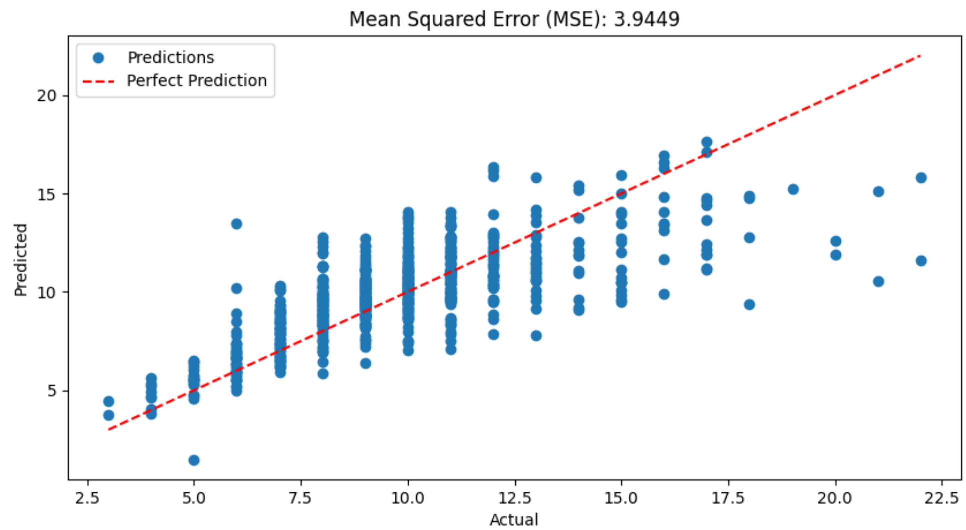

Figure 2: Mean Squared Error plot for Abalone dataset - Feed Forward Network

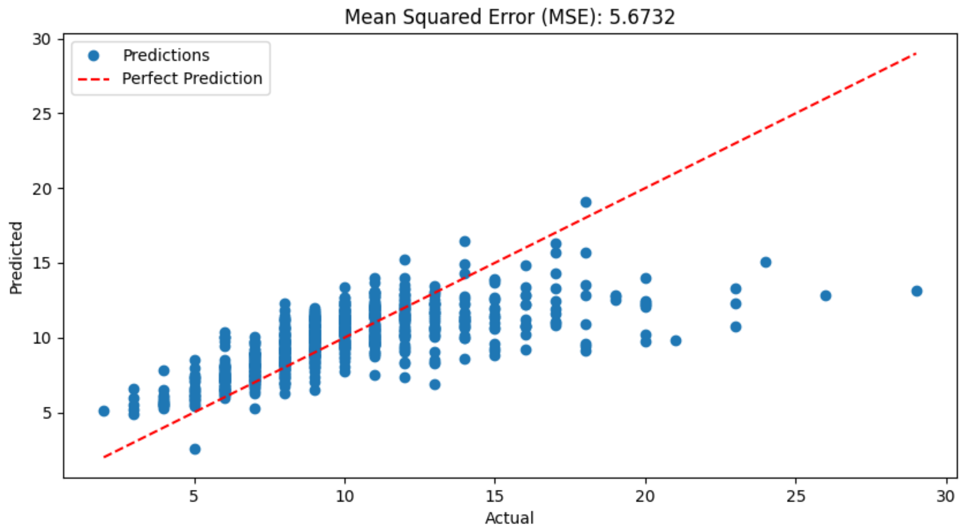

Figure 3: Mean Squared Error plot for Abalone dataset - Linear Regression

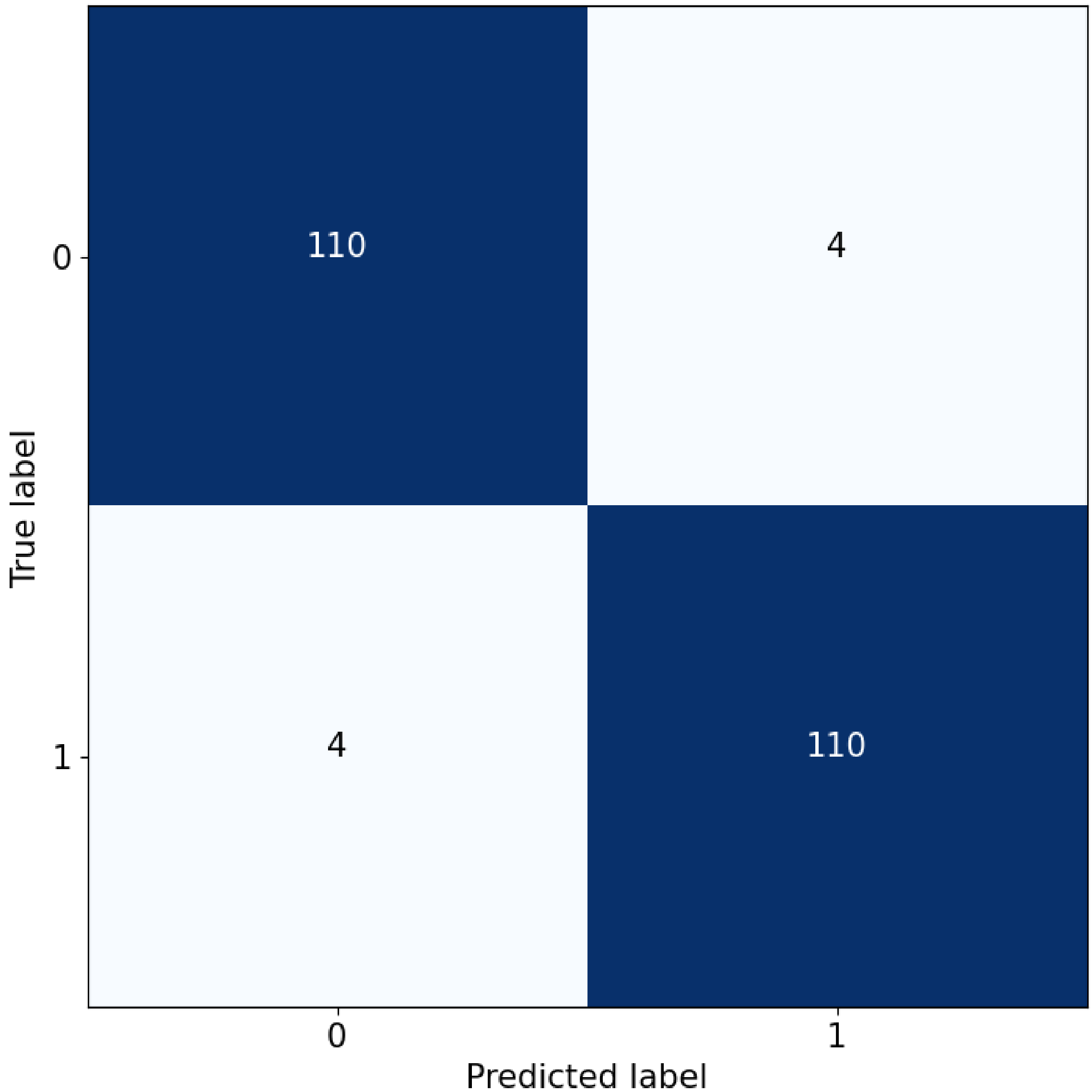

Figure 4: Mean Squared Error plot for Breast Cancer dataset - Combined Network

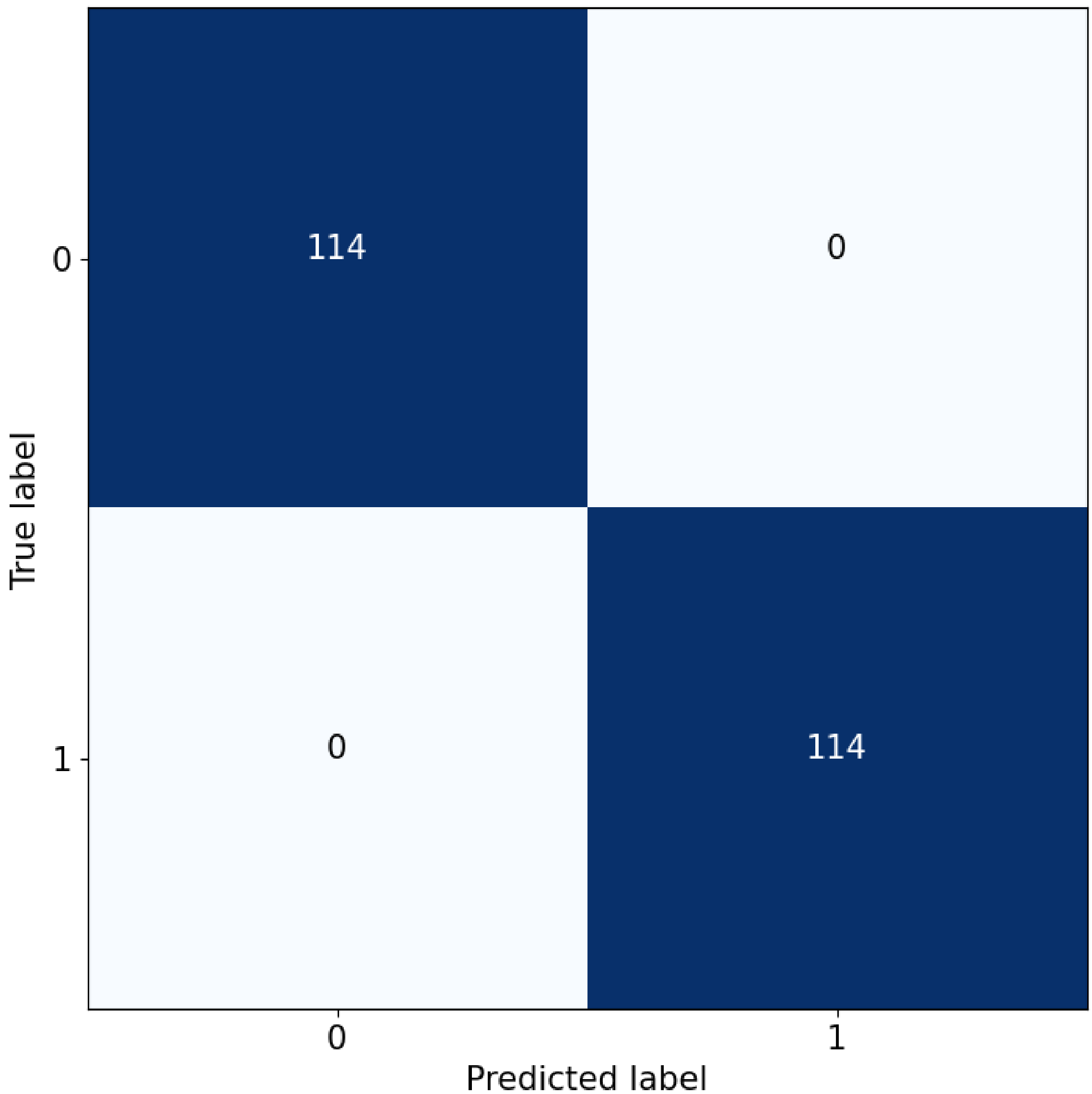

Figure 5: Mean Squared Error plot for Breast Cancer dataset - Feed Forward Network

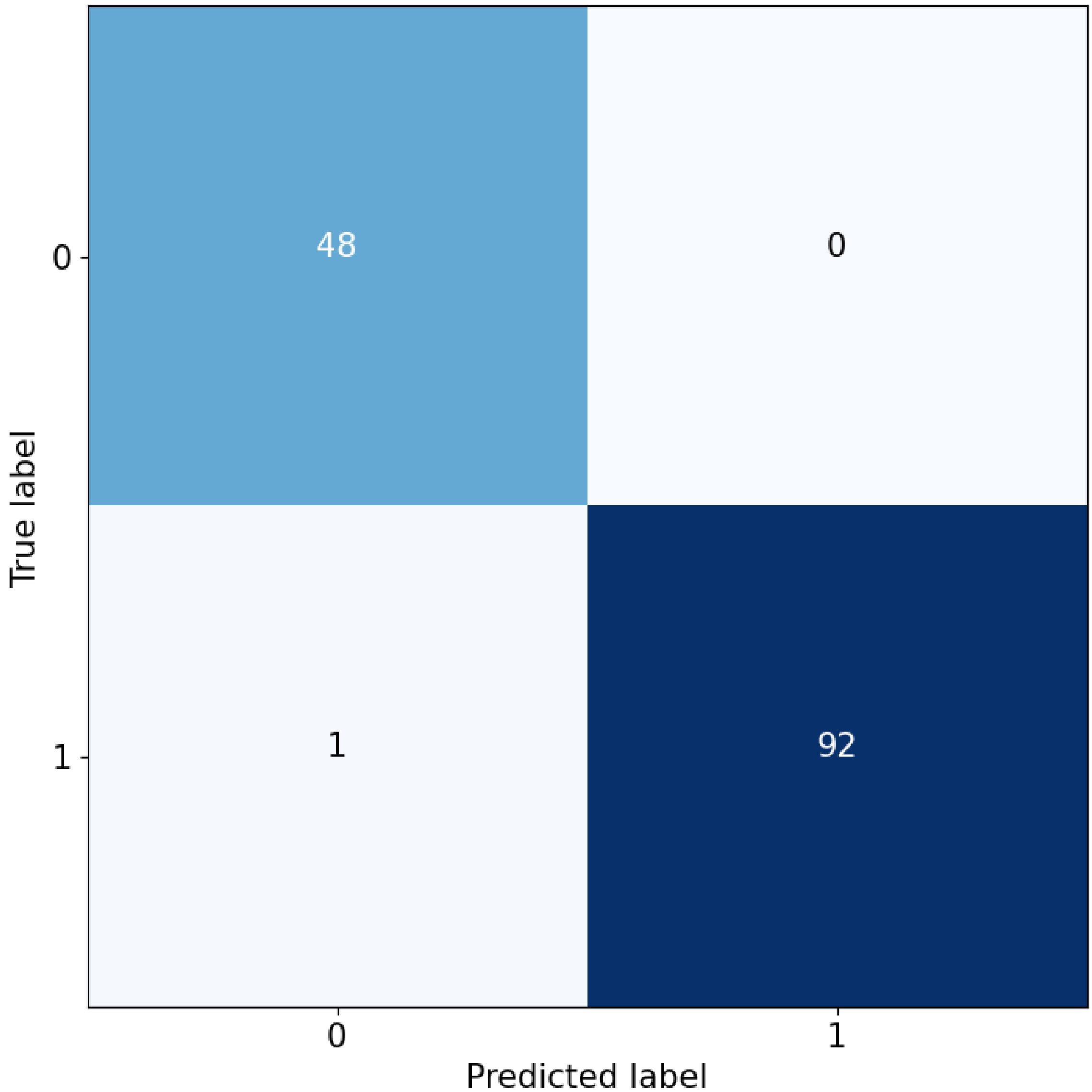

Figure 6: Mean Squared Error plot for Breast Cancer dataset - Logistic Regression

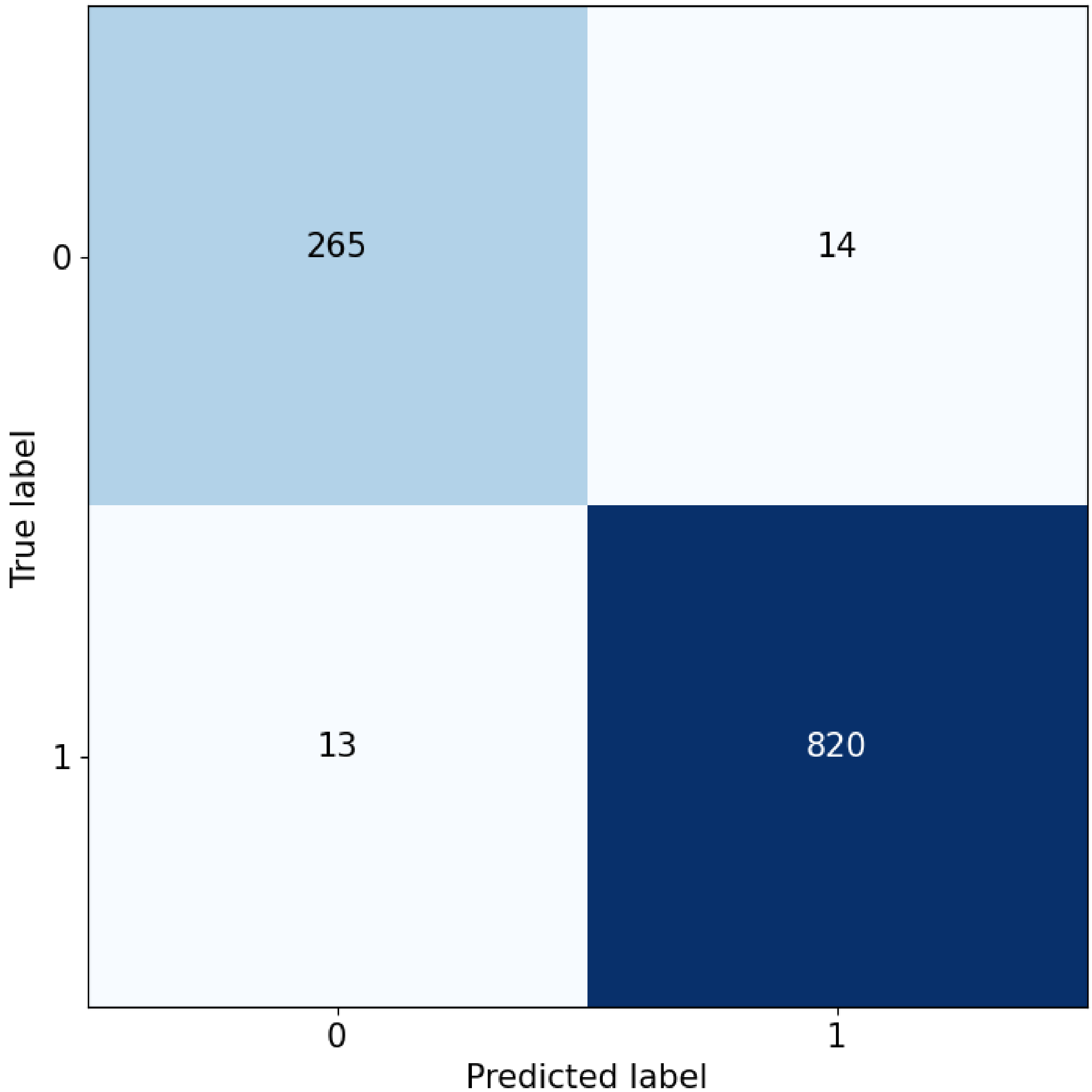

Figure 7: Mean Squared Error plot for Car Evaluation dataset - Combined Network

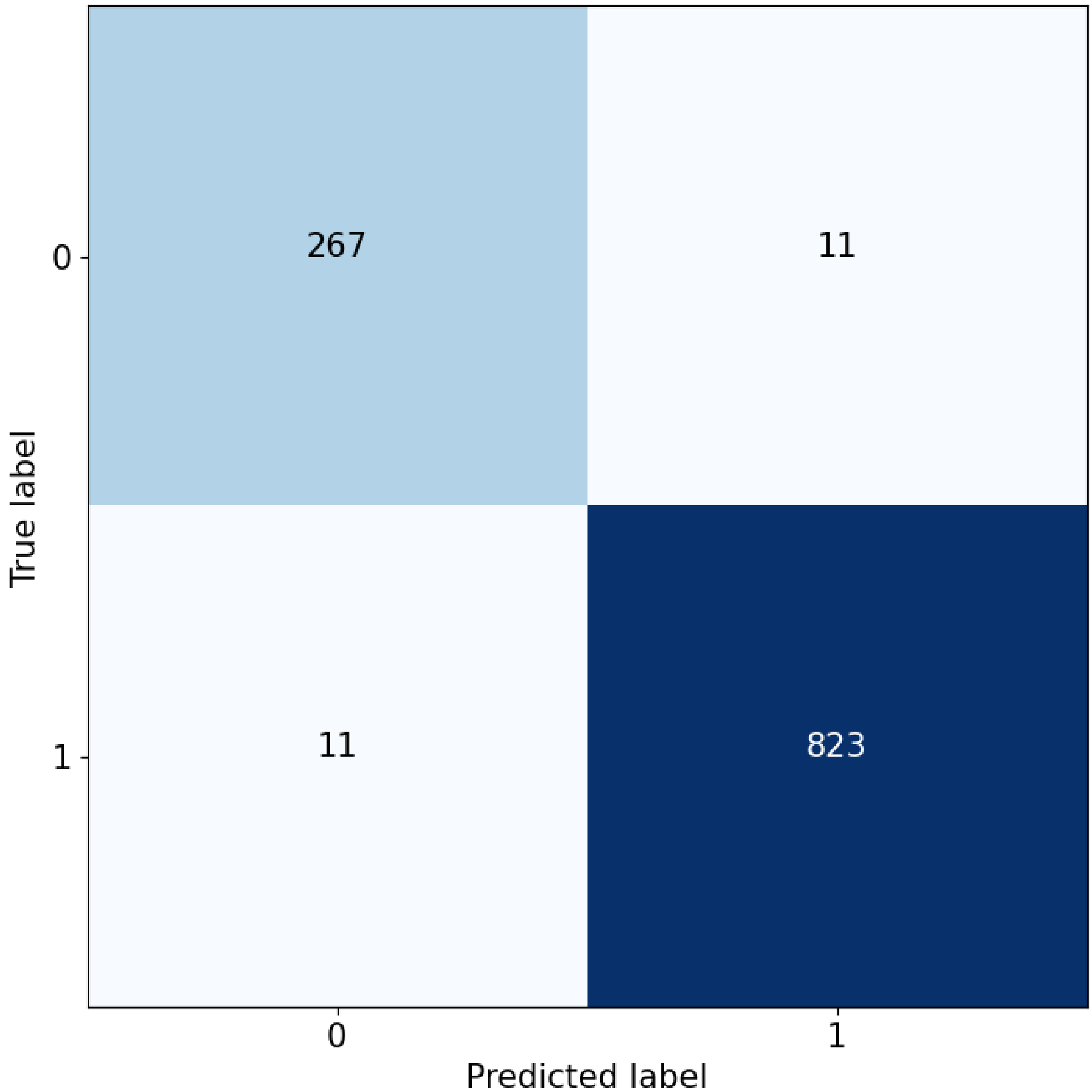

Figure 8: Mean Squared Error plot for Car Evaluation dataset - Feed Forward Network

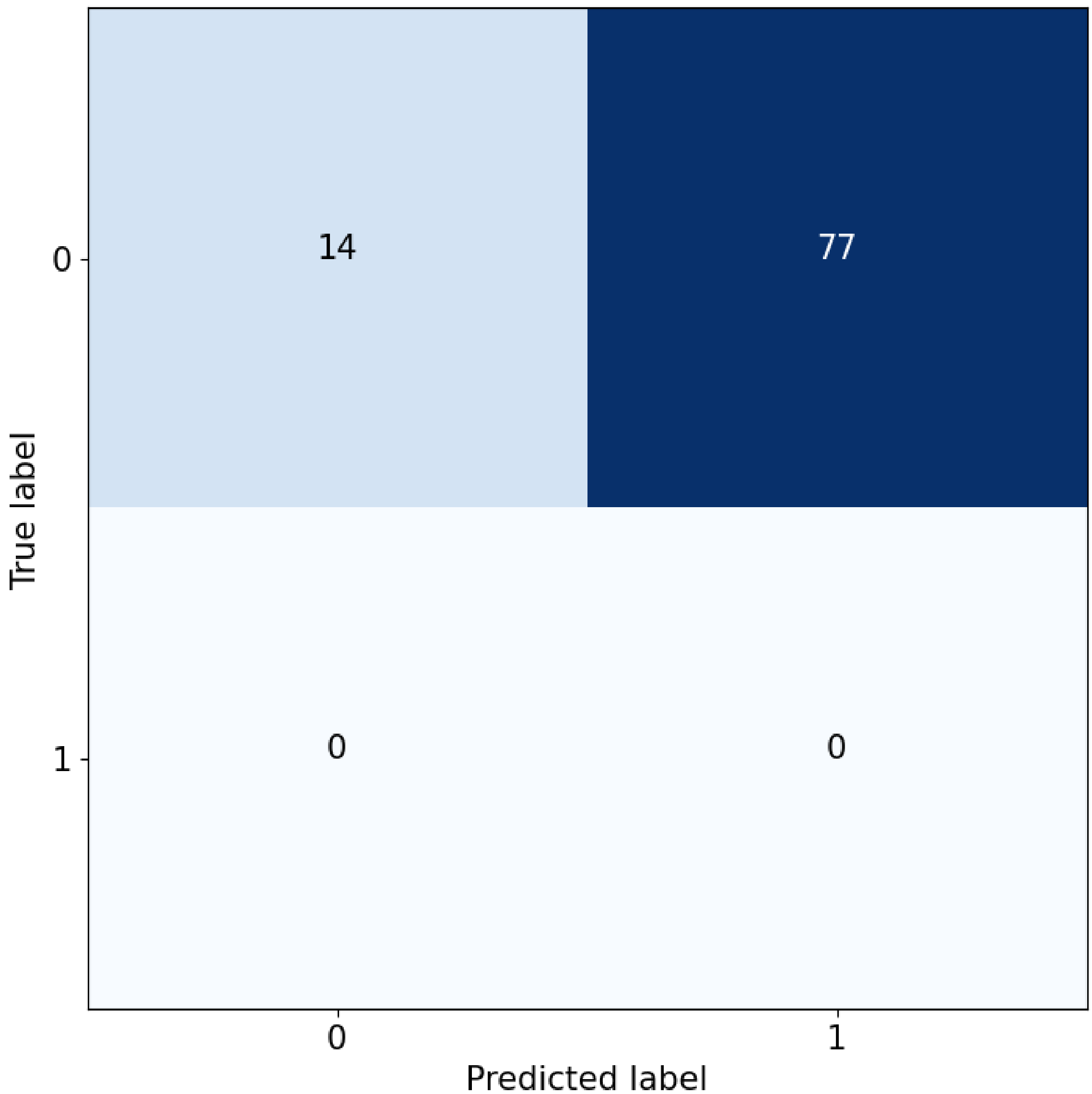

Figure 9: Mean Squared Error plot for Car Evaluation dataset - Logistic Regression

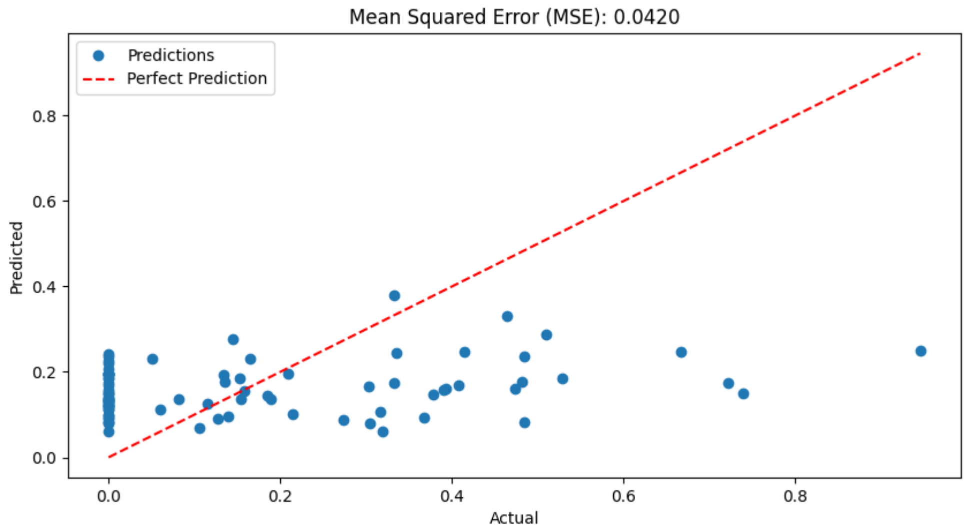

Figure 10: Mean Squared Error plot for Forest Fires dataset - Combined Network

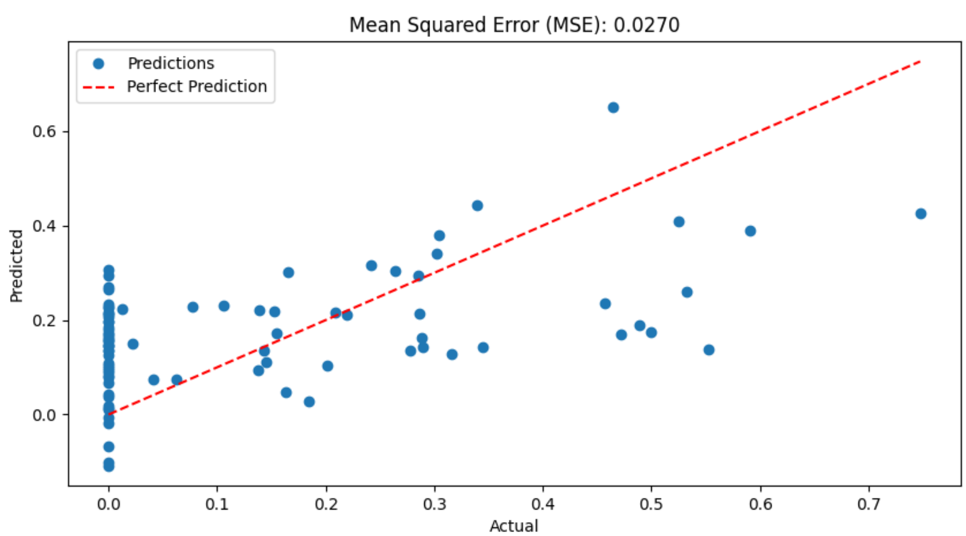

Figure 11: Mean Squared Error plot for Forest Fires dataset - Feed Forward Network

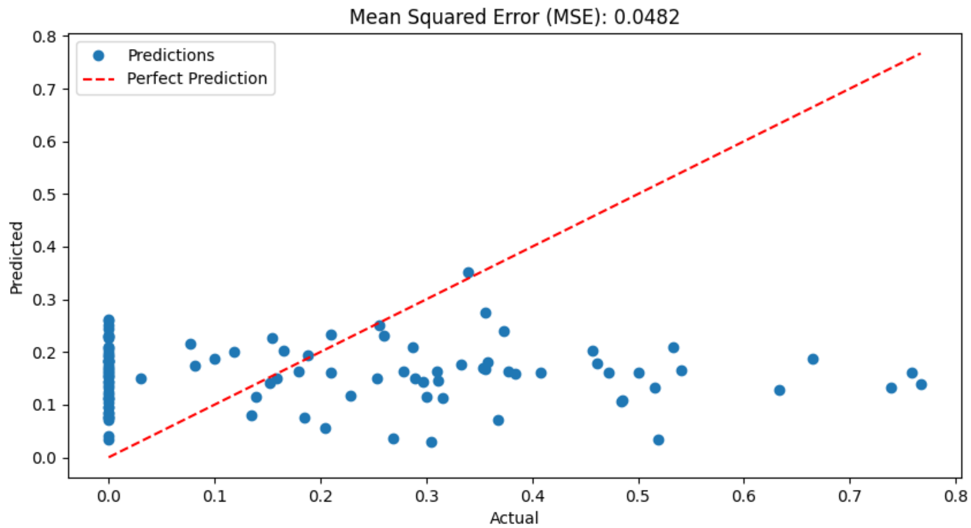

Figure 12: Mean Squared Error plot for Forest Fires dataset - Linear Regression

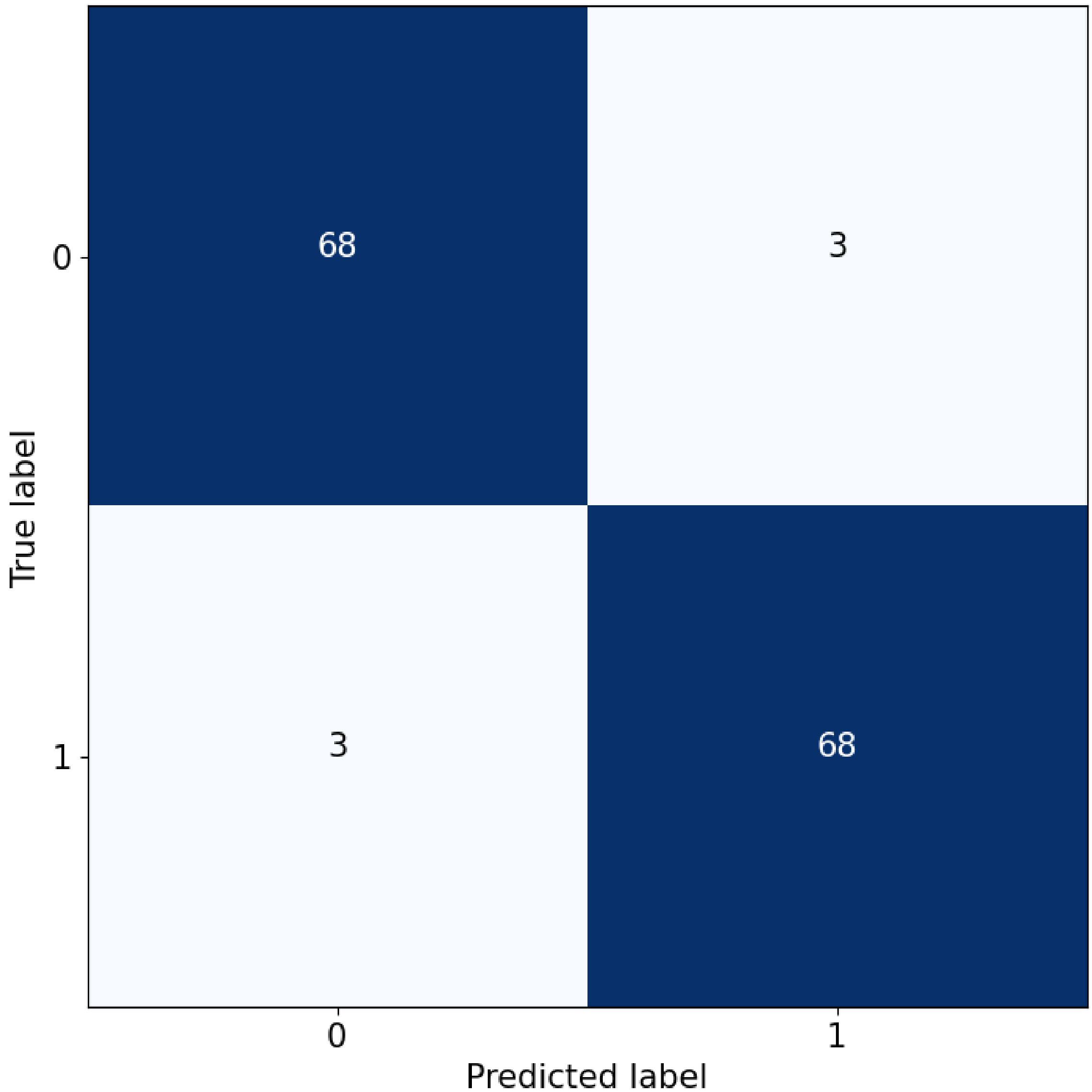

Figure 13: Mean Squared Error plot for House Votes 84 dataset - Combined Network

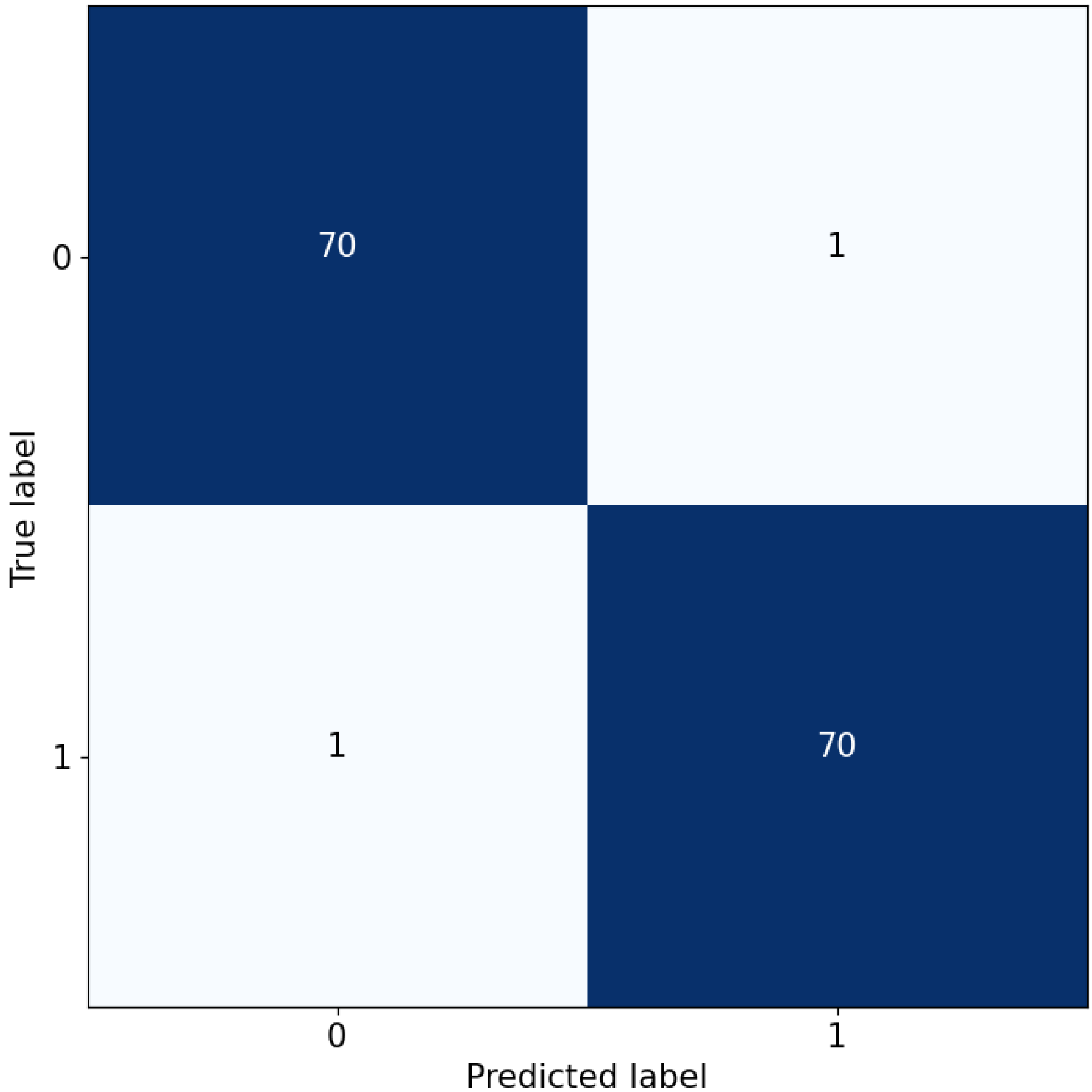

Figure 14: Mean Squared Error plot for House Votes 84 dataset - Feed Forward Network

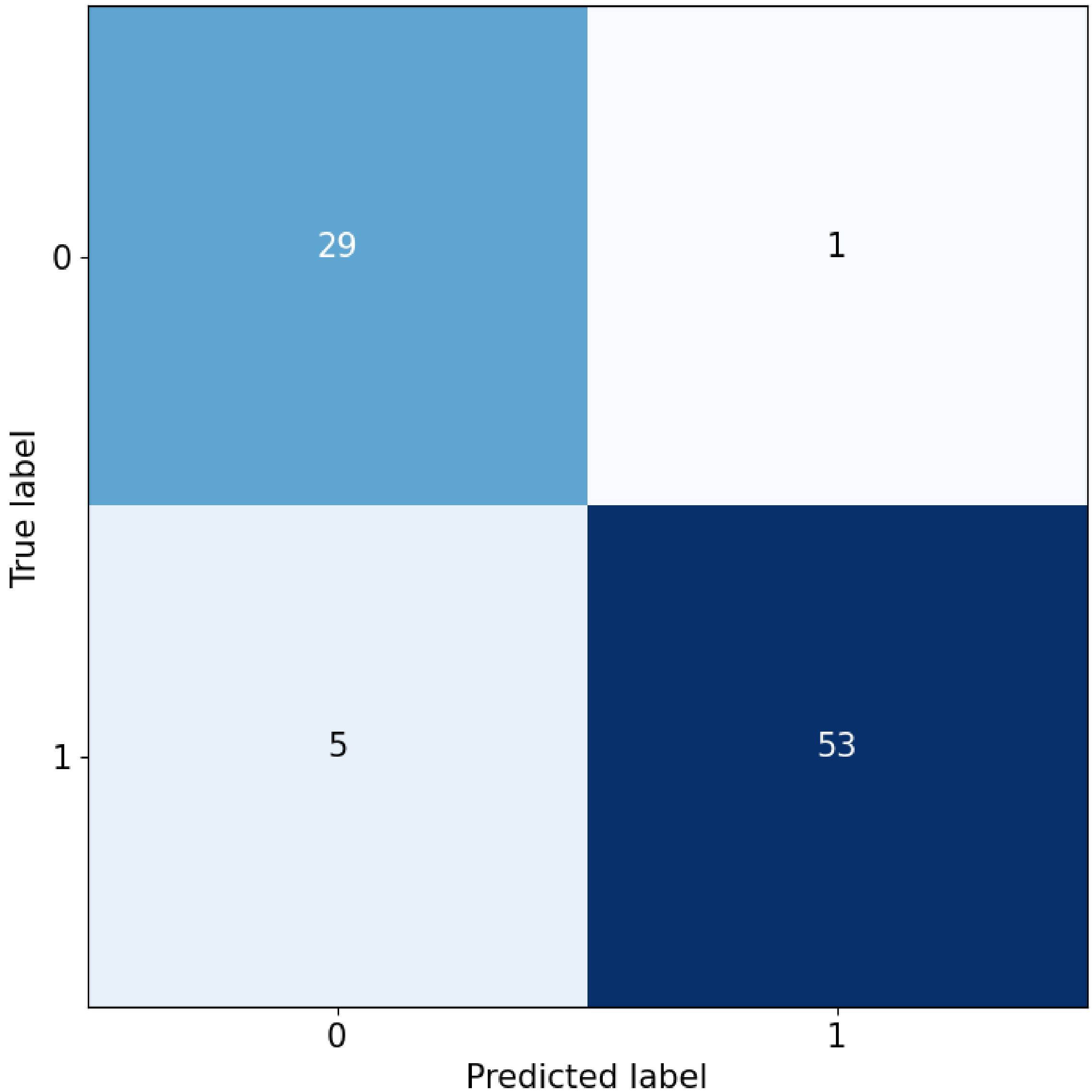

Figure 15: Mean Squared Error plot for House Votes 84 dataset - Logistic Regression

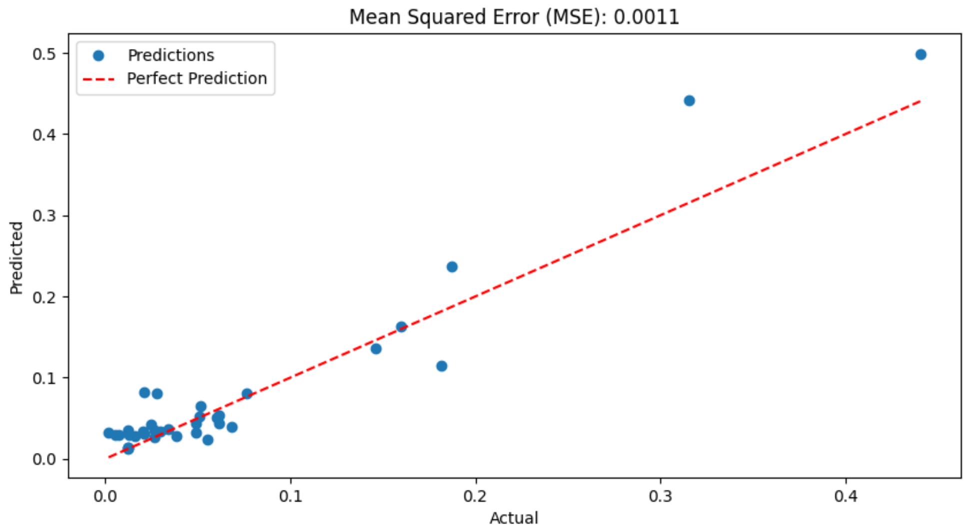

Figure 16: Mean Squared Error plot for Machine dataset - Combined Network

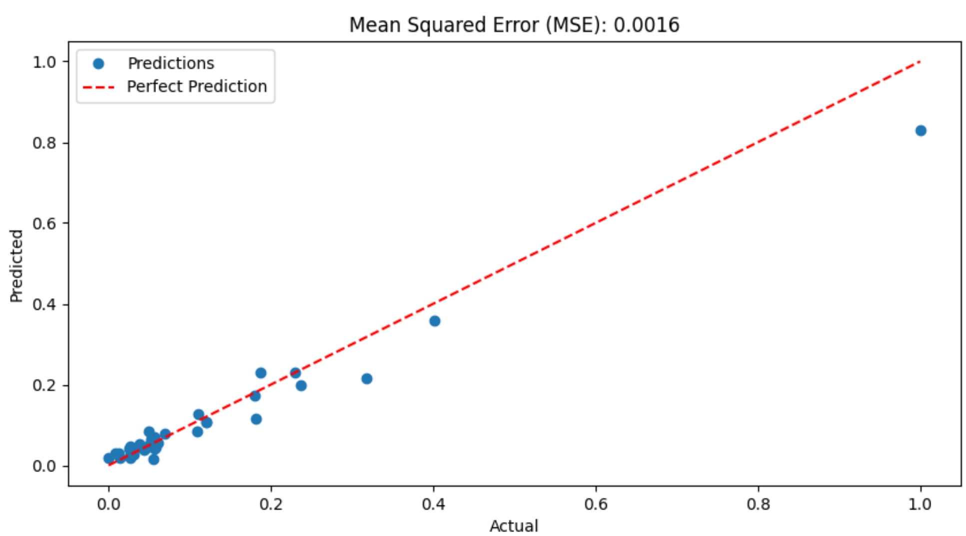

Figure 17: Mean Squared Error plot for Machine dataset - Feed Forward Network

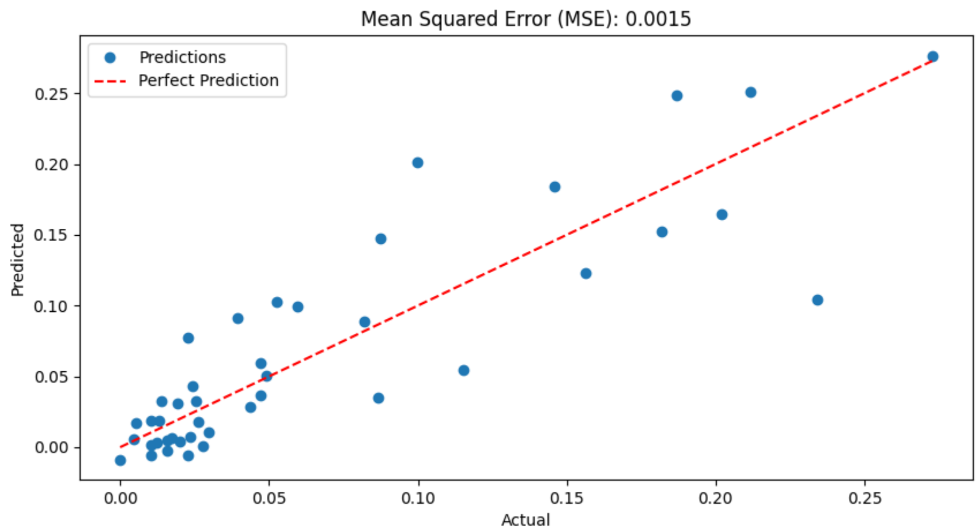

Figure 18: Mean Squared Error plot for Machine dataset - Linear Regression