KAUNAS UNIVERSITY OF TECHNOLOGY

INFORMATICS FACULTY

INTRODUCTION TO ARTIFICIAL INTELLIGENCE
DATA ANALYSIS LAB 3 WORK REPORT

Report author: Uysal Demirci

Lecturer: doc.dr. Germanas Budnikas

Kaunas, 2024

1. Initial Dataset Description

Selected Dataset: Guns

Link: https://vincentarelbundock.github.io/Rdatasets/csv/AER/Guns.csv

Description: Guns is a balanced panel of data on 50 US states, plus the District of Columbia (for a total of 51 states), by year for 1977–1999.

Format: A data frame containing 1,173 observations on 13 variables.

> ***state*** factor indicating state.
>
> ***year*** factor indicating year.
>
> ***violent*** violent crime rate (incidents per 100,000 members of the population).
>
> ***murder*** murder rate (incidents per 100,000).
>
> ***robbery*** robbery rate (incidents per 100,000).
>
> ***prisoners*** incarceration rate in the state in the previous year (sentenced prisoners per 100,000 residents; value for the previous year).
>
> ***afam*** percent of state population that is African-American, ages 10 to 64.
>
> ***cauc*** percent of state population that is Caucasian, ages 10 to 64.
>
> ***male*** percent of state population that is male, ages 10 to 29.
>
> ***population*** state population, in millions of people.
>
> ***income*** real per capita personal income in the state (US dollars).
>
> ***density*** population per square mile of land area, divided by 1,000.
>
> ***law*** factor. Does the state have a shall carry law in effect in that year?

```
RangeIndex: 1173 entries, 0 to 1172
Data columns (total 13 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   year        1173 non-null   int64
 1   violent     1173 non-null   float64
 2   murder      1173 non-null   float64
 3   robbery     1173 non-null   float64
 4   prisoners   1173 non-null   int64
 5   afam        1173 non-null   float64
 6   cauc        1173 non-null   float64
 7   male        1173 non-null   float64
 8   population  1173 non-null   float64
 9   income      1173 non-null   float64
 10  density     1173 non-null   float64
 11  state       1173 non-null   object
 12  law         1173 non-null   object
dtypes: float64(9), int64(2), object(2)
memory usage: 119.3+ KB
```

Fig.1: Datatypes of columns.

2. Dataset Rearrangements

The categorical variables 'state' and 'law' were transformed into numerical variables to be used in the model because the artificial neural network model can only be trained with numerical variables.

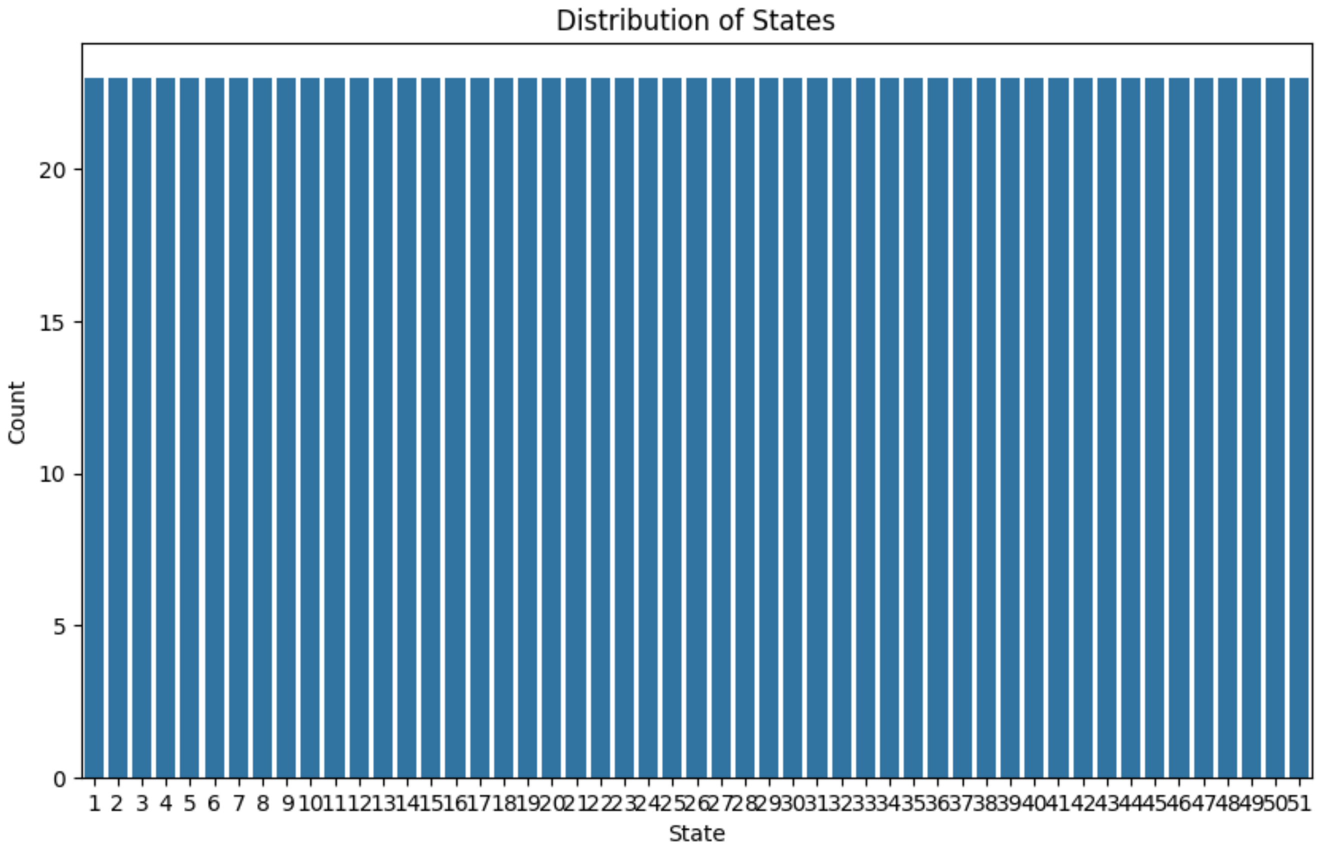

Fig.2: Distribution of States.

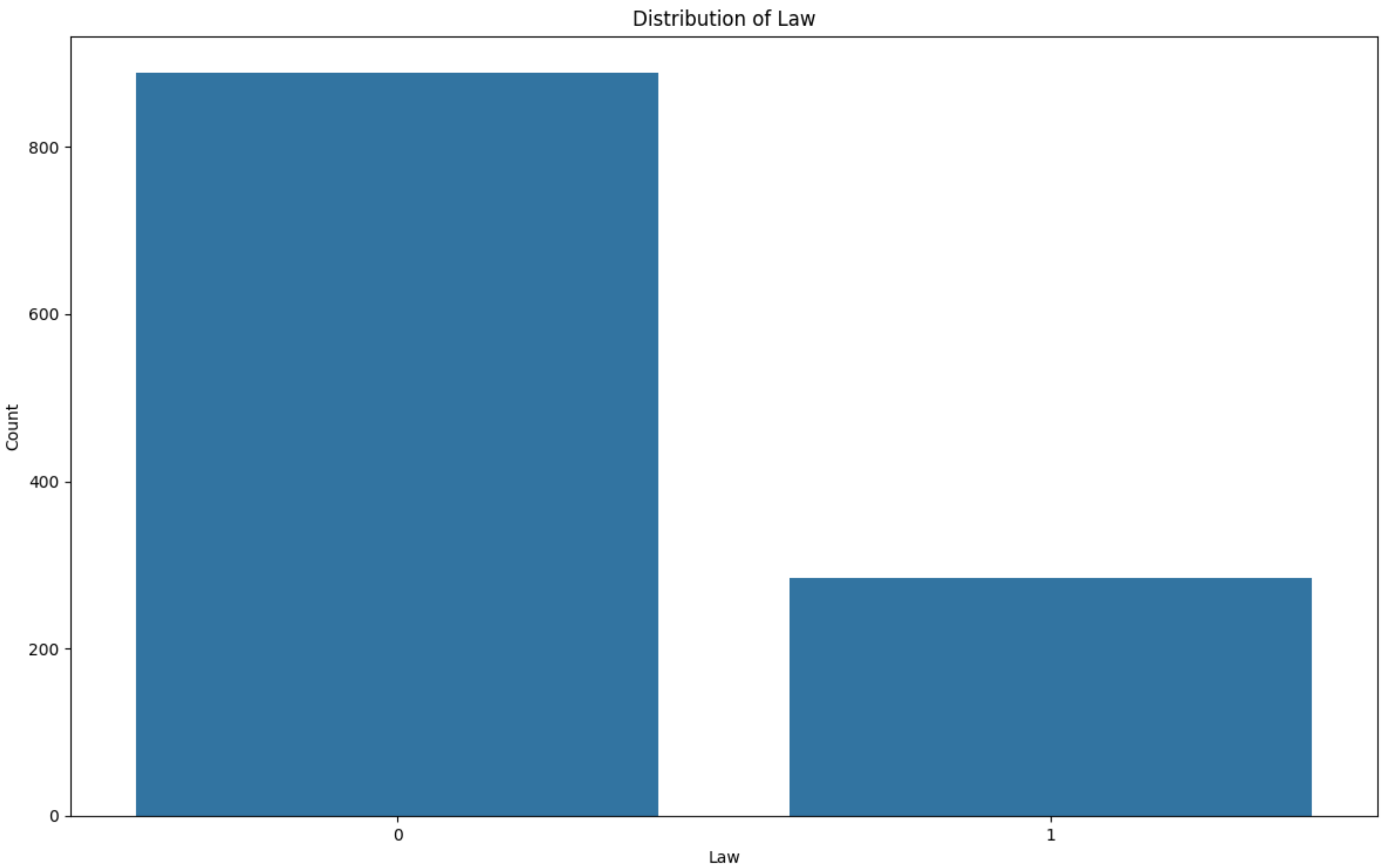

Fig.3: Distribution of Law.

3. Architecture of ANN

Input Layer: The input layer consists of 12 neurons corresponding to the 12 features in the dataset.

Hidden Layers:

First Hidden Layer: 6 neurons with ReLU activation function.

Second Hidden Layer: 6 neurons with ReLU activation function.

Output Layer: 1 neuron with Hyperbolic Tangent activation function.

Parameters:

Learning Rate: 0.0005

Activation Function: ReLU for hidden layers, Hyperbolic Tangent for the output layer

Batch Size: 10

Epochs: 100

This ANN architecture is designed to handle the classification task with the given dataset. The model uses the Adam optimizer with a learning rate of 0.0005 and binary cross-entropy loss function. The ReLU activation function is employed in the hidden layers to introduce non-linearity, while the Hyperbolic Tangent activation function in the output layer provides probabilities for binary classification.

4. Results of 10-Fold Cross Validation Experiments

```
Accuracy Scores at Each Fold:
Fold 1: 0.8829787234042553
Fold 2: 0.8617021276595744
Fold 3: 0.7659574468085106
Fold 4: 0.7659574468085106
Fold 5: 0.8297872340425532
Fold 6: 0.7659574468085106
Fold 7: 0.8191489361702128
Fold 8: 0.7659574468085106
Fold 9: 0.8709677419354839
Fold 10: 0.7634408602150538

Cost Function Values at Each Fold:
Fold 1: 0.2780149738991417
Fold 2: 0.28170358105262455
Fold 3: 3.7311966164510815
Fold 4: 3.7311966164510815
Fold 5: 0.3153614383195524
Fold 6: 0.5338227432197732
Fold 7: 0.30656096968645014
Fold 8: 3.7311966164510815
Fold 9: 0.3152852272006799
Fold 10: 3.771317008894542

Average Cost Function Value: 1.6995655791626008
Average Accuracy: 80.92%
```

Fig.4: Results of 10-fold cross validation

5. Measures In Order To Increase ANN Performance.

To improve the performance of the ANN model and achieve a better accuracy rate, the activation function used in the output layer was changed to the **Sigmoid** function.

```
Accuracy Scores at Each Fold:
Fold 1: 0.851063829787234
Fold 2: 0.8404255319148937
Fold 3: 0.8617021276595744
Fold 4: 0.8404255319148937
Fold 5: 0.776595744680851
Fold 6: 0.7446808510638298
Fold 7: 0.7978723404255319
Fold 8: 0.851063829787234
Fold 9: 0.8602150537634409
Fold 10: 0.8279569892473119

Cost Function Values at Each Fold:
Fold 1: 0.29221829517720077
Fold 2: 0.3781730931768926
Fold 3: 0.3526153907727188
Fold 4: 0.38044636760255823
Fold 5: 0.39608289095924404
Fold 6: 0.44660621663727246
Fold 7: 0.42293325425930245
Fold 8: 0.3095272999421088
Fold 9: 0.3314774677315812
Fold 10: 0.37434072168435867

Average Cost Function Value: 0.36844209979432374
Average Accuracy: 82.52%
```

Fig.5: Results of 10-fold cross validation after changing activation function.

After this change, the accuracy rate increased from 80.92% to 82.52%.

Since the increase in accuracy was not sufficient, the learning rate was adjusted in an attempt to achieve the desired accuracy. The learning rate, which was initially set at 0.0005, was changed to 0.01, and the process was repeated.

```
Accuracy Scores at Each Fold:
Fold 1: 0.9148936170212766
Fold 2: 0.8723404255319149
Fold 3: 0.8936170212765957
Fold 4: 0.851063829787234
Fold 5: 0.9361702127659575
Fold 6: 0.8191489361702128
Fold 7: 0.8404255319148937
Fold 8: 0.925531914893617
Fold 9: 0.9354838709677419
Fold 10: 0.9139784946236559

Cost Function Values at Each Fold:
Fold 1: 0.24179405223946365
Fold 2: 0.21926423985283452
Fold 3: 0.2909779972248499
Fold 4: 0.2947166964624757
Fold 5: 0.18021184350466105
Fold 6: 0.4233169653845
Fold 7: 0.270358967050825
Fold 8: 0.2741902708829665
Fold 9: 0.1398680828483176
Fold 10: 0.2905938254490567

Average Cost Function Value: 0.26252929408999504
Average Accuracy: 89.03%
```

Fig.6: Results of 10-fold cross validation after changing learning rate.

As a result of these operations, the accuracy reached 89.03%. More than the desired 5% increase was achieved.

6. Conclusions

In conclusion, through the iterative process of building, refining, and optimizing an artificial neural network (ANN) model for prediction or classification tasks, significant improvements in performance were achieved. Beginning with the transformation of categorical variables into numerical ones to enable model training, we designed an ANN architecture with an input layer comprising 12 neurons, followed by two hidden layers each with 6 neurons and ReLU activation functions. The output layer utilized the Sigmoid activation function to enhance classification accuracy.

By fine-tuning hyperparameters such as the learning rate and activation functions, we successfully increased the model's accuracy from an initial 80.92% to 89.03%, surpassing our target of a 5% improvement. Specifically, adjusting the learning rate from 0.0005 to 0.01 yielded substantial performance gains.

These results underscore the importance of meticulous experimentation and parameter tuning in optimizing machine learning models. They also demonstrate the effectiveness of artificial neural networks in handling complex datasets and achieving high prediction accuracy. Moving forward, further exploration of model architectures and hyperparameters will be crucial not only to enhance performance but also to gain new insights from the data.