Name: Raven Jacinto Course and Section: CPE019 - CPE32S3 Date of Submission: April 3, 2024 Instructor: Engr. Roman Richard Activity: Midterm Quiz 1: Interactive Lab (Neural Networks)

Instructions:

```
Use this provided dataset: PhiUSIIL Phishing URL (Website) - UCI
Machine Learning Repository
```

Links to an external site. Perform:

```
Task 1: Exploratory Data Analysis (Cleaning + Prepping the dataset)
Task 2: Data modelling using ANN
```

Importing needed libraries

```
pip install ucimlrepo
```

```
Requirement already satisfied: ucimlrepo in
/usr/local/lib/python3.10/dist-packages (0.0.6)
```

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler

## Import Keras objects for Deep Learning

from sklearn.metrics import confusion_matrix, precision_recall_curve,
roc_auc_score, roc_curve, accuracy_score
from keras.models  import Sequential
from keras.layers import Input, Dense, Flatten, Dropout,
BatchNormalization
from keras.optimizers import Adam, SGD, RMSprop
```

Loading the dataset

```
from ucimlrepo import fetch_ucirepo

# fetch dataset
phiusiil_phishing_url_website = fetch_ucirepo(id=967)

# data (as pandas dataframes)
X = phiusiil_phishing_url_website.data.features
y = phiusiil_phishing_url_website.data.targets
```

```
# metadata
print(phiusiil_phishing_url_website.metadata)

# variable information
print(phiusiil_phishing_url_website.variables)
```

```
{'uci_id': 967, 'name': 'PhiUSIIL Phishing URL (Website)',
'repository_url':
'https://archive.ics.uci.edu/dataset/967/phiusiil+phishing+url+dataset
', 'data_url':
'https://archive.ics.uci.edu/static/public/967/data.csv', 'abstract':
'PhiUSIIL Phishing URL Dataset is a substantial dataset comprising
134,850 legitimate and 100,945 phishing URLs. Most of the URLs we
analyzed, while constructing the dataset, are the latest URLs.
Features are extracted from the source code of the webpage and URL.
Features such as CharContinuationRate, URLTitleMatchScore,
URLCharProb, and TLDLegitimateProb are derived from existing
features.', 'area': 'Computer Science', 'tasks': ['Classification'],
'characteristics': ['Tabular'], 'num_instances': 235795,
'num_features': 54, 'feature_types': ['Real', 'Categorical',
'Integer'], 'demographics': [], 'target_col': ['label'], 'index_col':
None, 'has_missing_values': 'no', 'missing_values_symbol': None,
'year_of_dataset_creation': 2024, 'last_updated': 'Mon Mar 18 2024',
'dataset_doi': 'https://doi.org/10.1016/j.cose.2023.103545',
'creators': ['Arvind Prasad', 'Shalini Chandra'], 'intro_paper':
{'title': 'PhiUSIIL: A diverse security profile empowered phishing URL
detection framework based on similarity index and incremental
learning', 'authors': 'Arvind Prasad and Shalini Chandra',
'published_in': 'Computers & Security', 'year': 2024, 'url':
'https://doi.org/10.1016/j.cose.2023.103545', 'doi': None},
'additional_info': {'summary': None, 'purpose': None, 'funded_by':
None, 'instances_represent': 'URLs and their corresponding webpages',
'recommended_data_splits': None, 'sensitive_data': None,
'preprocessing_description': None, 'variable_info': 'Column "FILENAME"
can be ignored.', 'citation': 'Prasad, A., & Chandra, S. (2023).
PhiUSIIL: A diverse security profile empowered phishing URL detection
framework based on similarity index and incremental learning.
Computers & Security, 103545. doi:
https://doi.org/10.1016/j.cose.2023.103545'}}
                          name     role         type demographic
description  \
0                     FILENAME    Other  Categorical        None
None
1                          URL  Feature  Categorical        None
None
2                    URLLength  Feature      Integer        None
None
3                       Domain  Feature  Categorical        None
None
4                 DomainLength  Feature      Integer        None
```

```
None
5                 IsDomainIP  Feature      Integer        None
None
6                        TLD  Feature  Categorical        None
None
7         URLSimilarityIndex  Feature      Integer        None
None
8       CharContinuationRate  Feature      Integer        None
None
9          TLDLegitimateProb  Feature   Continuous        None
None
10               URLCharProb  Feature   Continuous        None
None
11                 TLDLength  Feature      Integer        None
None
12             NoOfSubDomain  Feature      Integer        None
None
13            HasObfuscation  Feature      Integer        None
None
14        NoOfObfuscatedChar  Feature      Integer        None
None
15          ObfuscationRatio  Feature      Integer        None
None
16          NoOfLettersInURL  Feature      Integer        None
None
17          LetterRatioInURL  Feature   Continuous        None
None
18           NoOfDegitsInURL  Feature      Integer        None
None
19           DegitRatioInURL  Feature      Integer        None
None
20           NoOfEqualsInURL  Feature      Integer        None
None
21            NoOfQMarkInURL  Feature      Integer        None
None
22        NoOfAmpersandInURL  Feature      Integer        None
None
23  NoOfOtherSpecialCharsInURL  Feature      Integer        None
None
24     SpacialCharRatioInURL  Feature   Continuous        None
None
25                   IsHTTPS  Feature      Integer        None
None
26                LineOfCode  Feature      Integer        None
None
27         LargestLineLength  Feature      Integer        None
None
28                  HasTitle  Feature      Integer        None
None
```

```
29                       Title  Feature  Categorical         None
None
30       DomainTitleMatchScore  Feature      Integer         None
None
31          URLTitleMatchScore  Feature      Integer         None
None
32                  HasFavicon  Feature      Integer         None
None
33                      Robots  Feature      Integer         None
None
34                IsResponsive  Feature      Integer         None
None
35             NoOfURLRedirect  Feature      Integer         None
None
36            NoOfSelfRedirect  Feature      Integer         None
None
37              HasDescription  Feature      Integer         None
None
38                   NoOfPopup  Feature      Integer         None
None
39                  NoOfiFrame  Feature      Integer         None
None
40       HasExternalFormSubmit  Feature      Integer         None
None
41                HasSocialNet  Feature      Integer         None
None
42             HasSubmitButton  Feature      Integer         None
None
43             HasHiddenFields  Feature      Integer         None
None
44            HasPasswordField  Feature      Integer         None
None
45                        Bank  Feature      Integer         None
None
46                         Pay  Feature      Integer         None
None
47                      Crypto  Feature      Integer         None
None
48            HasCopyrightInfo  Feature      Integer         None
None
49                   NoOfImage  Feature      Integer         None
None
50                     NoOfCSS  Feature      Integer         None
None
51                      NoOfJS  Feature      Integer         None
None
52                 NoOfSelfRef  Feature      Integer         None
None
53                NoOfEmptyRef  Feature      Integer         None
```

```
None
54               NoOfExternalRef  Feature      Integer        None
None
55                         label   Target      Integer        None
None

   units missing_values
0   None             no
1   None             no
2   None             no
3   None             no
4   None             no
5   None             no
6   None             no
7   None             no
8   None             no
9   None             no
10  None             no
11  None             no
12  None             no
13  None             no
14  None             no
15  None             no
16  None             no
17  None             no
18  None             no
19  None             no
20  None             no
21  None             no
22  None             no
23  None             no
24  None             no
25  None             no
26  None             no
27  None             no
28  None             no
29  None             no
30  None             no
31  None             no
32  None             no
33  None             no
34  None             no
35  None             no
36  None             no
37  None             no
38  None             no
39  None             no
40  None             no
41  None             no
```

```
42   None                 no
43   None                 no
44   None                 no
45   None                 no
46   None                 no
47   None                 no
48   None                 no
49   None                 no
50   None                 no
51   None                 no
52   None                 no
53   None                 no
54   None                 no
55   None                 no
```

Upon checking, there lots of integer data that is why I will turn other categorical features into numbers using astype. I will also drop some tables which I do think that is not needed for this model. I will transoform values in some columns because it may overpower others and for it to have good generalization.

Cleaning of data / Preprocessing

```
X.head()
```

{"type":"dataframe","variable_name":"X"}

```
y.head()
```

{"type":"dataframe","variable_name":"y"}

```
X.shape
```

```
(235795, 54)
```

```
y.shape
```

```
(235795, 1)
```

```
#Dropping unwanted columns
X.drop('URL', axis = 1, inplace = True)
X.drop('Domain', axis = 1, inplace = True)
X.drop('Title', axis = 1, inplace = True)
```

```
X.shape
```

```
<ipython-input-39-a1ffef5a001d>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
```

```
  X.drop('URL', axis = 1, inplace = True)
<ipython-input-39-a1ffef5a001d>:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
  X.drop('Domain', axis = 1, inplace = True)
<ipython-input-39-a1ffef5a001d>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
  X.drop('Title', axis = 1, inplace = True)

(235795, 51)
```

```
#Checking if unwanted columns is dropped successfully
X.head()
```

```
{"type":"dataframe","variable_name":"X"}
```

```
#Converting categorical types into integer

X['TLD'] = X['TLD'].astype('category')
X['TLD'] = X['TLD'].cat.codes
```

```
<ipython-input-47-2b686cfb721f>:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
  X['TLD'] = X['TLD'].astype('category')
<ipython-input-47-2b686cfb721f>:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
  X['TLD'] = X['TLD'].cat.codes
```

```
X['TLD']
```

```
0         231
1         254
2         647
3         231
```

```
4         503
         ...
235790    231
235791    647
235792    157
235793    258
235794    231
Name: TLD, Length: 235795, dtype: int16
```

```
X.head()
```

```
{"type":"dataframe","variable_name":"X"}
```

-The data consist of all integer values now.

-Data does not have any categorical values now.

-Unwanted columns were dropped.

Data Splitting 75% training and 25% testing

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.25, random_state=11111)
```

```
np.mean(y), np.mean(1-y)
```

```
/usr/local/lib/python3.10/dist-packages/numpy/core/
fromnumeric.py:3502: FutureWarning: In a future version,
DataFrame.mean(axis=None) will return a scalar mean over the entire
DataFrame. To retain the old behavior, use 'frame.mean(axis=0)' or
just 'frame.mean()'
  return mean(axis=axis, dtype=dtype, out=out, **kwargs)
/usr/local/lib/python3.10/dist-packages/numpy/core/fromnumeric.py:3502
: FutureWarning: In a future version, DataFrame.mean(axis=None) will
return a scalar mean over the entire DataFrame. To retain the old
behavior, use 'frame.mean(axis=0)' or just 'frame.mean()'
  return mean(axis=axis, dtype=dtype, out=out, **kwargs)
```

```
(label    0.571895
 dtype: float64,
 label    0.428105
 dtype: float64)
```

```
# Scaling the value for better model performance

normalizer = StandardScaler()
X_train_norm = normalizer.fit_transform(X_train)
X_test_norm = normalizer.transform(X_test)
```

Neural Network

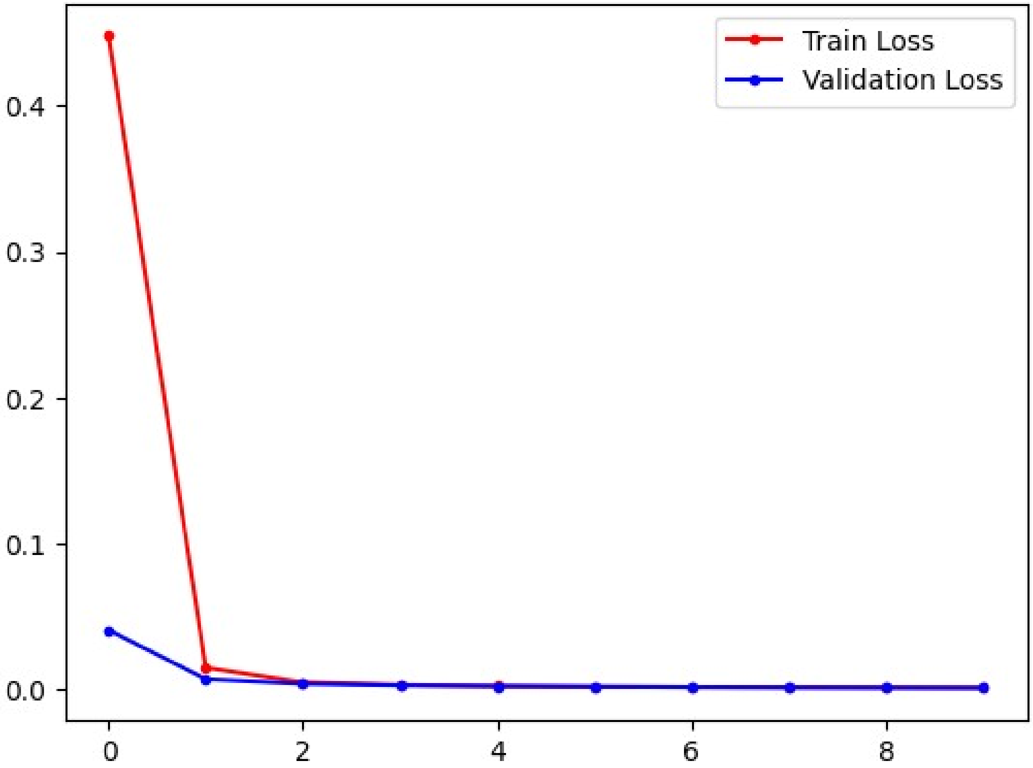
```
#Artificial Neural Network

model  = Sequential([
    Dense(32, input_shape=(51,), activation="softmax"),
    Dense(16, input_shape=(51,), activation="relu"),
    Dense(8, input_shape=(51,), activation="relu"),
    Dense(1, activation="sigmoid")
])

#Cmodel.summary()

Model: "sequential_2"
_________________________________________________________________
 Layer (type)                Output Shape              Param #   
=================================================================
 dense_16 (Dense)            (None, 32)                1664      
                                                                 
 dense_17 (Dense)            (None, 16)                528       
                                                                 
 dense_18 (Dense)            (None, 8)                 136       
                                                                 
 dense_19 (Dense)            (None, 1)                 9         
                                                                 
=================================================================
Total params: 2337 (9.13 KB)
Trainable params: 2337 (9.13 KB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________

# Compilation of the model
model.compile(SGD(lr = 0.001), "binary_crossentropy",
metrics=["accuracy"])
run_hist_1 = model.fit(X_train_norm, y_train,
validation_data=(X_test_norm, y_test), epochs=10)

WARNING:absl:`lr` is deprecated in Keras optimizer, please use
`learning_rate` or use the legacy optimizer,
e.g.,tf.keras.optimizers.legacy.SGD.

Epoch 1/10
5527/5527 [==============================] - 14s 2ms/step - loss:
0.4478 - accuracy: 0.7494 - val_loss: 0.0407 - val_accuracy: 0.9924
Epoch 2/10
5527/5527 [==============================] - 14s 3ms/step - loss:
0.0151 - accuracy: 0.9968 - val_loss: 0.0073 - val_accuracy: 0.9984
Epoch 3/10
5527/5527 [==============================] - 15s 3ms/step - loss:
0.0052 - accuracy: 0.9988 - val_loss: 0.0041 - val_accuracy: 0.9991
Epoch 4/10
```

```
5527/5527 [==============================] - 17s 3ms/step - loss:
0.0034 - accuracy: 0.9992 - val_loss: 0.0030 - val_accuracy: 0.9993
Epoch 5/10
5527/5527 [==============================] - 18s 3ms/step - loss:
0.0026 - accuracy: 0.9994 - val_loss: 0.0024 - val_accuracy: 0.9994
Epoch 6/10
5527/5527 [==============================] - 15s 3ms/step - loss:
0.0021 - accuracy: 0.9995 - val_loss: 0.0021 - val_accuracy: 0.9995
Epoch 7/10
5527/5527 [==============================] - 18s 3ms/step - loss:
0.0018 - accuracy: 0.9996 - val_loss: 0.0018 - val_accuracy: 0.9996
Epoch 8/10
5527/5527 [==============================] - 14s 3ms/step - loss:
0.0016 - accuracy: 0.9996 - val_loss: 0.0016 - val_accuracy: 0.9996
Epoch 9/10
5527/5527 [==============================] - 14s 3ms/step - loss:
0.0014 - accuracy: 0.9997 - val_loss: 0.0015 - val_accuracy: 0.9996
Epoch 10/10
5527/5527 [==============================] - 14s 3ms/step - loss:
0.0012 - accuracy: 0.9997 - val_loss: 0.0014 - val_accuracy: 0.9996
```

Plotting of the model performance (training loss and validation loss)

```
fig, ax = plt.subplots()
ax.plot(run_hist_1.history["loss"],'r', marker='.', label="Train
Loss")
ax.plot(run_hist_1.history["val_loss"],'b', marker='.',
label="Validation Loss")
ax.legend()
```

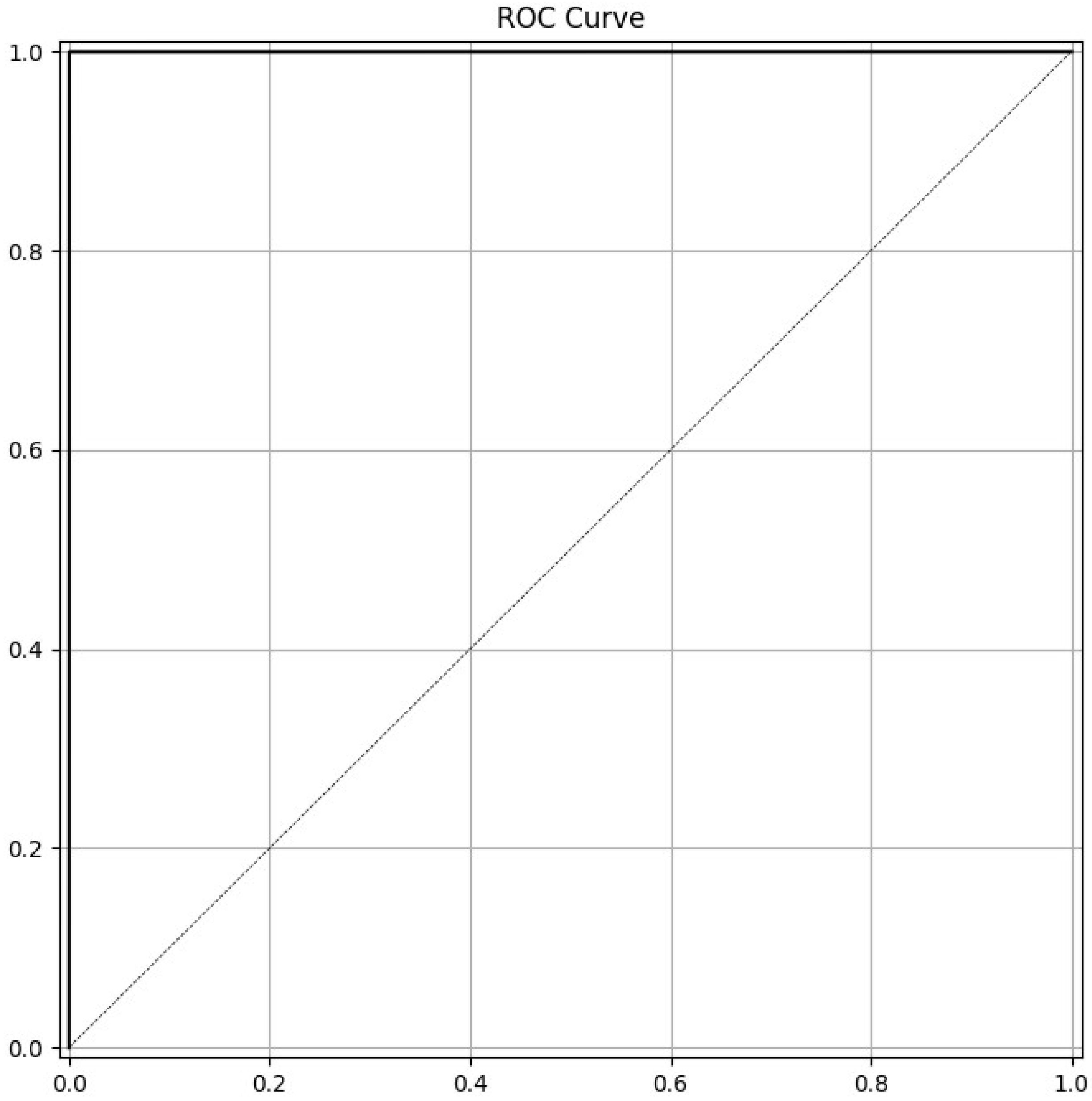
```
<matplotlib.legend.Legend at 0x7fb8361e2e30>
```



Evaluation: Since the plot shows that the train loss is high at first but on the second iteration it matches the loss in the validation and in the the third iteration, they are both close to each other, meaning training and validation loss performed the same. The closer the graph (validation and training loss) together, the more it is good for the model. We can also assume that the accuracy graph is close to each other too. This plot presents that the model has good generalization, has a balanced complexity, and performs stable training.

```
def plot_roc(y_test, y_pred, model_name):
    fpr, tpr, thr = roc_curve(y_test, y_pred)
    fig, ax = plt.subplots(figsize=(8, 8))
    ax.plot(fpr, tpr, 'k-')
    ax.plot([0, 1], [0, 1], 'k--', linewidth=.5)  # roc curve for
random model
    ax.grid(True)
    ax.set(title='ROC Curve'.format(model_name),
           xlim=[-0.01, 1.01], ylim=[-0.01, 1.01])

y_pred_prob_nn_1 = model.predict(X_test_norm)
y_pred_class_nn_1 = np.argmax(y_pred_prob_nn_1, axis=1)

print('accuracy is
{:.3f}'.format(accuracy_score(y_test,y_pred_class_nn_1)))
print('roc-auc is
{:.3f}'.format(roc_auc_score(y_test,y_pred_prob_nn_1)))

plot_roc(y_test, y_pred_prob_nn_1, 'NN')
```

```
1843/1843 [==============================] - 7s 4ms/step
accuracy is 0.428
roc-auc is 1.000
```



Evaluation: Using the ROC curve, we can identify that the model performed outstanding since, the ROC-AUC is 1.000, closer to top left meaning the model is learns a lot from the data that we have been dealing with and the model is performing very well.

Conclusion

In conclusion, I do think that I managed to complete the task or instructions in this activity. Which are performing data analysis, cleaning the dataset or preprocessing. Model all the data using artificial neural networks and plot its results. I enjoyed doing this activity and I refreshed my knowledge in neural network.