Malicious URL Detection

a) Dataset Selection:

While conducting our experiments, a crucial aspect involved in the selection of an appropriate dataset. The dataset chosen played a pivotal role in shaping the outcomes and conclusions of our study.

Dataset Name: "**malicious_phish.csv**"

Description: The Dataset comprises a total of 6,51,191 URLs, each categorized different types namely phishing, malware, defacement and benign.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 651191 entries, 0 to 651190
Data columns (total 3 columns):
 #   Column  Non-Null Count   Dtype 
---  ------  --------------   ----- 
 0   url     651191 non-null  object
 1   type    651191 non-null  object
 2   tld     175910 non-null  object
dtypes: object(3)
memory usage: 14.9+ MB
```

	url	type
0	br-icloud.com.br	phishing
1	mp3raid.com/music/krizz_kaliko.html	benign
2	bopsecrets.org/rexroth/cr/1.htm	benign
3	http://www.garage-pirenne.be/index.php?option=com_	defacement
4	http://adventure-nicaragua.net/index.php?option=co	defacement
5	http://buzzfil.net/m/show-art/ils-etaient-loin-de-	benign
6	espn.go.com/nba/player/_/id/3457/brandon-rush	benign
7	yourbittorrent.com/?q=anthony-hamilton-soulife	benign
8	http://www.pashminaonline.com/pure-pashminas	defacement
9	allmusic.com/album/crazy-from-the-heat-r16990	benign
10	corporationwiki.com/Ohio/Columbus/frank-s-benson-P	benign
11	http://www.ikenmijnkunst.nl/index.php/expositie s/e	defacement
12	myspace.com/video/vid/30602581	benign
13	http://www.lebensmittel-ueberwachung.de/index.php/	defacement
14	http://www.szabadmunkaero.hu/cimoldal.html?start=1	defacement
15	http://larcadelcarnevale.com/catalogo/palloncini	defacement
16	quickfacts.census.gov/qfd/maps/iowa_map.html	benign
17	nugget.ca/ArticleDisplay.aspx?archive=true&e=11609	benign
18	uk.linkedin.com/pub/steve-rubenstein/8/718/755	benign
19	http://www.vnic.co/khach-hang.html	defacement

b) Task 1: Sample Runs of feature creation

For the creation of features, we initially preprocessed the dataset by checking if the dataset contains null values.

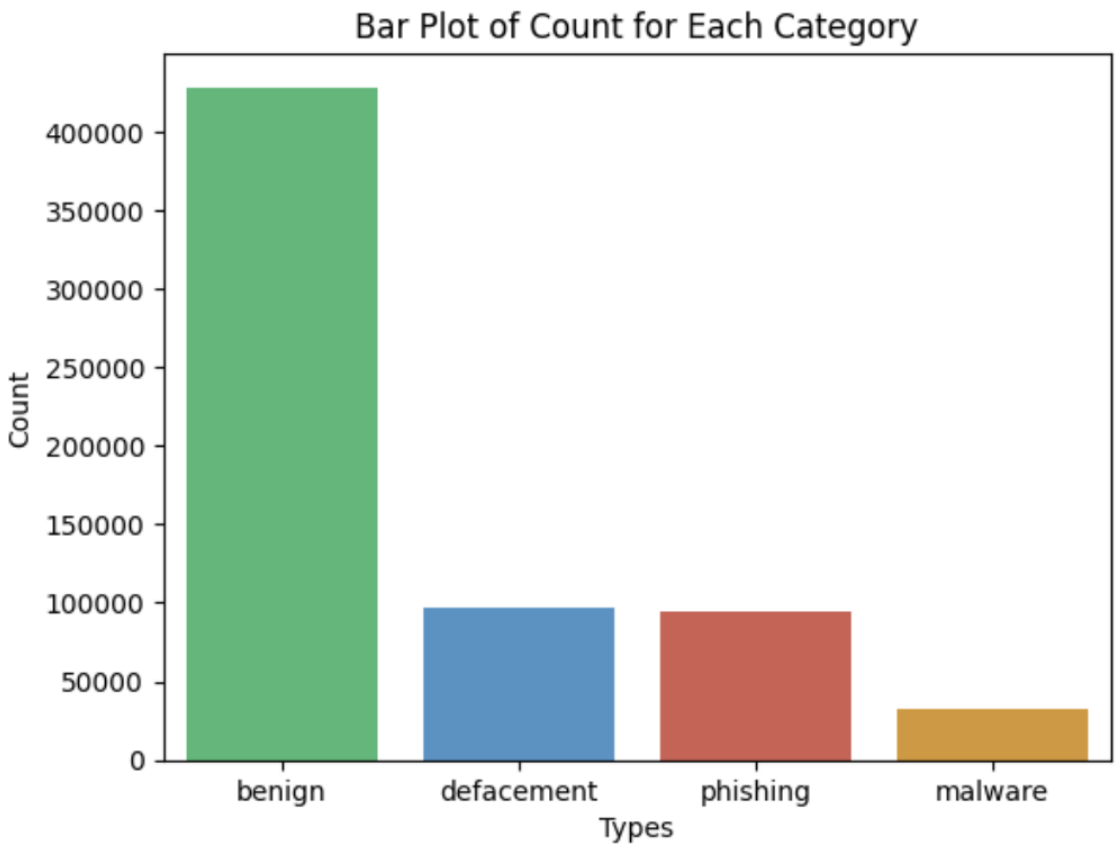
```
[ ]  # Check for missing values in each column of the DataFrame and sum them up
     data.isnull().sum()
```

```
url          0
type         0
tld     475281
dtype: int64
```

> By extracting the types of data that contains in the dataset.

```
[ ]  # Extracting the index of the 'count' Series, which represents categories or types
     x=count.index
     x
```

```
Index(['benign', 'defacement', 'phishing', 'malware'], dtype='object')
```



> Removing the “www.” From the URLs columns, so it would be easier for the machine to learn and train.

```
[ ]  # Remove 'www.' from the 'url' column using the replace method and a regular expression
     data['url'] = data['url'].replace('www.', '', regex=True)
     data
```

	url	type	tld
0	br-icloud.com.br	phishing	None
1	mp3raid.com/music/krizz_kaliko.html	benign	None
2	bopsecrets.org/rexroth/cr/1.htm	benign	None
3	http://garage-pirenne.be/index.php?option=com_...	defacement	garage-pirenne.be
4	http://adventure-nicaragua.net/index.php?optio...	defacement	adventure-nicaragua.net
...	...	...	...
651186	xbox360.ign.com/objects/850/850402.html	phishing	None
651187	games.teamxbox.com/xbox-360/1860/Dead-Space/	phishing	None
651188	gamespot.com/xbox360/action/deadspace/	phishing	None
651189	en.wikipedia.org/wiki/Dead_Space_(video_game)	phishing	None
651190	angelfire.com/goth/devilmaycrytonite/	phishing	None

651191 rows × 3 columns

> In the Further step categorizing these types from 0 to 3, starting benign: 0, defacement: 1, phishing: 2 , malware: 3

	url	type	tld	Category
0	br-icloud.com.br	phishing	None	2
1	mp3raid.com/music/krizz_kaliko.html	benign	None	0
2	bopsecrets.org/rexroth/cr/1.htm	benign	None	0
3	http://garage-pirenne.be/index.php?option=com_...	defacement	garage-pirenne.be	1
4	http://adventure-nicaragua.net/index.php?optio...	defacement	adventure-nicaragua.net	1
5	http://buzzfil.net/m/show-art/ils-etaient-loin...	benign	buzzfil.net	0

> Creation of TFIDF Vectorizer

```
                                                 url        type  \
0                                  br-icloud.com.br    phishing
1              mp3raid.com/music/krizz_kaliko.html      benign
2                  bopsecrets.org/rexroth/cr/1.htm      benign
3  http://garage-pirenne.be/index.php?option=com_...  defacement
4  http://adventure-nicaragua.net/index.php?optio...  defacement
5  http://buzzfil.net/m/show-art/ils-etaient-loin...      benign
6      espn.go.com/nba/player/_/id/3457/brandon-rush      benign
7     yourbittorrent.com/?q=anthony-hamilton-soulife      benign
8          http://pashminaonline.com/pure-pashminas  defacement
9      allmusic.com/album/crazy-from-the-heat-r16990      benign

                      tld  Category  000webhostapp   01   02   03   04   05  \
0                    None         2            0.0  0.0  0.0  0.0  0.0  0.0
1                    None         0            0.0  0.0  0.0  0.0  0.0  0.0
2                    None         0            0.0  0.0  0.0  0.0  0.0  0.0
3       garage-pirenne.be         1            0.0  0.0  0.0  0.0  0.0  0.0
4  adventure-nicaragua.net         1            0.0  0.0  0.0  0.0  0.0  0.0
5              buzzfil.net         0            0.0  0.0  0.0  0.0  0.0  0.0
6                    None         0            0.0  0.0  0.0  0.0  0.0  0.0
7                    None         0            0.0  0.0  0.0  0.0  0.0  0.0
8      pashminaonline.com         1            0.0  0.0  0.0  0.0  0.0  0.0
9                    None         0            0.0  0.0  0.0  0.0  0.0  0.0

   ...  wikipedia   wn  wordpress  world   wp   ws  yahoo  year  youtube   za
0  ...        0.0  0.0        0.0    0.0  0.0  0.0    0.0   0.0      0.0  0.0
1  ...        0.0  0.0        0.0    0.0  0.0  0.0    0.0   0.0      0.0  0.0
2  ...        0.0  0.0        0.0    0.0  0.0  0.0    0.0   0.0      0.0  0.0
3  ...        0.0  0.0        0.0    0.0  0.0  0.0    0.0   0.0      0.0  0.0
4  ...        0.0  0.0        0.0    0.0  0.0  0.0    0.0   0.0      0.0  0.0
5  ...        0.0  0.0        0.0    0.0  0.0  0.0    0.0   0.0      0.0  0.0
6  ...        0.0  0.0        0.0    0.0  0.0  0.0    0.0   0.0      0.0  0.0
7  ...        0.0  0.0        0.0    0.0  0.0  0.0    0.0   0.0      0.0  0.0
8  ...        0.0  0.0        0.0    0.0  0.0  0.0    0.0   0.0      0.0  0.0
9  ...        0.0  0.0        0.0    0.0  0.0  0.0    0.0   0.0      0.0  0.0

[10 rows x 504 columns]
```

	url	type	lem_url	15	3457	70	adventure
0	br-icloud.	phishing	br-icloud.	0	0	0	0
1	mp3raid.co	benign	mp3raid.co	0	0	0	0
2	bopsecrets	benign	bopsecrets	0	0	0	0
3	http://www	defacement	http://www	0.28807	0	0.28807	0
4	http://adv	defacement	http://adv	0	0	0	0.311168
5	http://buz	benign	http://buz	0	0	0	0
6	espn.go.co	benign	espn.go.co	0	0.377672	0	0
7	yourbittor	benign	yourbittor	0	0	0	0
8	http://www	defacement	http://www	0	0	0	0
9	allmusic.c	benign	allmusic.c	0	0	0	0

> Whether a URL has an IP Address or not:

	url	type	tld	Category	having_ip_address
0	br-icloud.com.br	phishing	None	2	0
1	mp3raid.com/music/krizz_kaliko.html	benign	None	0	0
2	bopsecrets.org/rexroth/cr/1.htm	benign	None	0	0
3	http://garage-pirenne.be/index.php?option=com_...	defacement	garage-pirenne.be	1	0
4	http://adventure-nicaragua.net/index.php?optio...	defacement	adventure-nicaragua.net	1	0

```
[ ]  # Display the counts of unique values in the 'having_ip_address' column
     data['having_ip_address'].value_counts()
```

```
0    638703
1     12488
Name: having_ip_address, dtype: int64
```

> The Number of dots in a URL, A URL with many dots is more likely to be a bad one.

	url	type	tld	Category	having_ip_address	num_dots	is_bad_url
0	br-icloud.com.br	phishing	None	2	0	2	False
1	mp3raid.com/music/krizz_kaliko.html	benign	None	0	0	2	False
2	bopsecrets.org/rexroth/cr/1.htm	benign	None	0	0	2	False
3	http://garage-pirenne.be/index.php?option=com_...	defacement	garage-pirenne.be	1	0	2	False
4	http://adventure-nicaragua.net/index.php?optio...	defacement	adventure-nicaragua.net	1	0	2	False

> The length of a URL. Long URLs are more likely to be bad ones.

	url	type	tld	Category	having_ip_address	num_dots	is_bad_url	url_len
0	br-icloud.com.br	phishing	None	2	0	2	False	16
1	mp3raid.com/music/krizz_kaliko.html	benign	None	0	0	2	False	35
2	bopsecrets.org/rexroth/cr/1.htm	benign	None	0	0	2	False	31
3	http://garage-pirenne.be/index.php?option=com_...	defacement	garage-pirenne.be	1	0	2	False	84
4	http://adventure-nicaragua.net/index.php?optio...	defacement	adventure-nicaragua.net	1	0	2	True	235

> Age of the domain name, is acquired by installing packages named '***whois'***

	url	type	tld	Category	having_ip_address	num_dots	is_bad_url	url_len	domain_age
0	br-icloud.com.br	phishing	None	2	0	2	True	16	0.0
1	mp3raid.com/music/krizz_kaliko.html	benign	None	0	0	2	False	35	24.0
2	bopsecrets.org/rexroth/cr/1.htm	benign	None	0	0	2	False	31	24.0
3	http://garage-pirenne.be/index.php?option=com_...	defacement	garage-pirenne.be	1	0	2	False	84	NaN
4	http://adventure-nicaragua.net/index.php?optio...	defacement	adventure-nicaragua.net	1	0	2	False	235	NaN

> Whether a URL has a redirection script. The redirection may direct users to bad websites.

	url	type	tld	Category	having_ip_address	num_dots	is_bad_url	url_len	has_redirection_script
0	br-icloud.com.br	phishing	None	2	0	2	False	16	None
1	mp3raid.com/music/krizz_kaliko.html	benign	None	0	0	2	False	35	None
2	bopsecrets.org/rexroth/cr/1.htm	benign	None	0	0	2	False	31	None
3	http://garage-pirenne.be/index.php?option=com_...	defacement	garage-pirenne.be	1	0	2	False	84	None
4	http://adventure-nicaragua.net/index.php?optio...	defacement	adventure-nicaragua.net	1	0	2	True	235	None

>> Whether URL contains JavaScript.

	url	type	tld	Category	having_ip_address	num_dots	is_bad_url	url_len	has_redirection_script	contains_javascript
0	br-icloud.com.br	phishing	None	2	0	2	False	16	None	None
1	mp3raid.com/music/krizz_kaliko.html	benign	None	0	0	2	False	35	None	None
2	bopsecrets.org/rexroth/cr/1.htm	benign	None	0	0	2	False	31	None	None
3	http://garage-pirenne.be/index.php?option=com_...	defacement	garage-pirenne.be	1	0	2	False	84	None	None
4	http://adventure-nicaragua.net/index.php?optio...	defacement	adventure-nicaragua.net	1	0	2	True	235	None	None

Throughout this process, our approach to feature creation was not solely guided by existing literature; we also introduced novel ideas based on our understanding of the domain. For example, the inclusion of domain age, URLs contains JavaScript and redirection scripts or not. For the above-mentioned features and representation of tables, I came up with new ideas to represent the way the output look by using the tabulate packages.

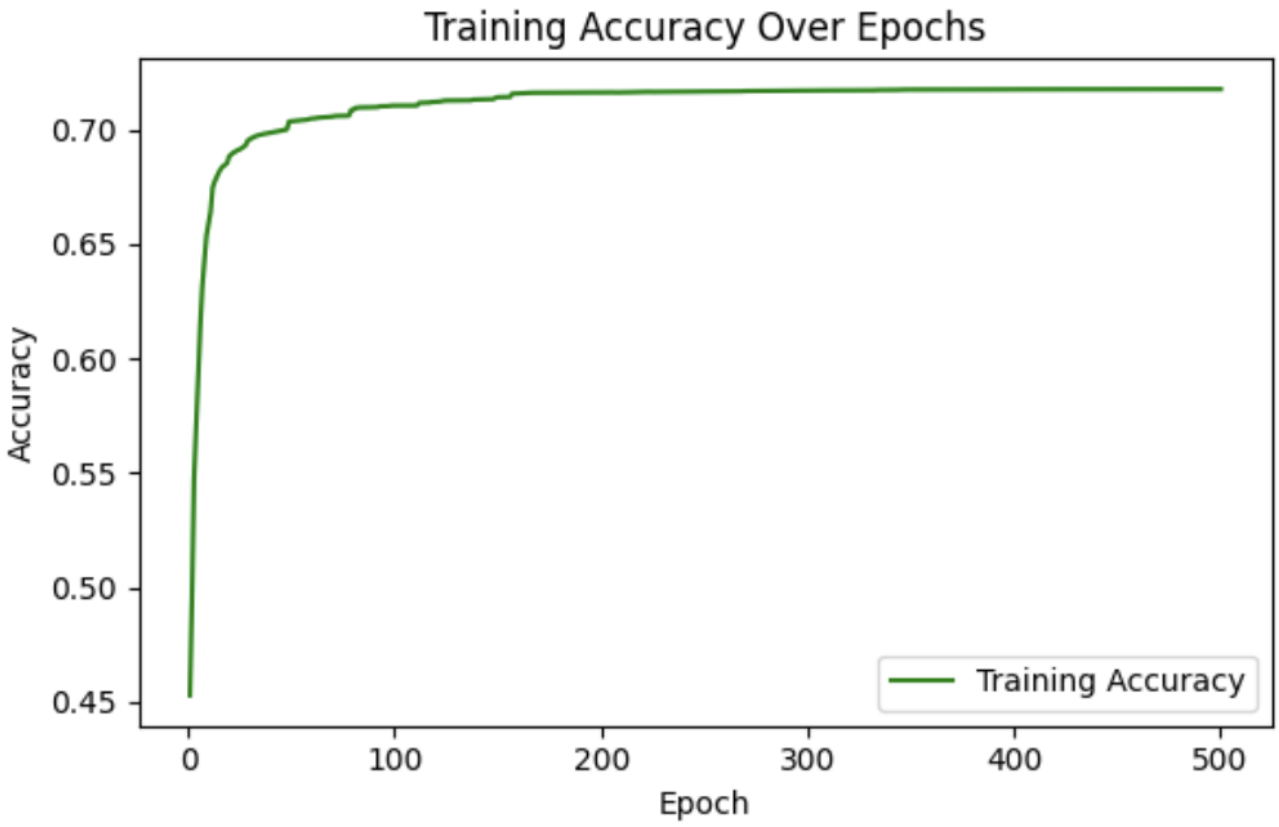
c) Task 2: Implementing Logistic Regression in PyTorch

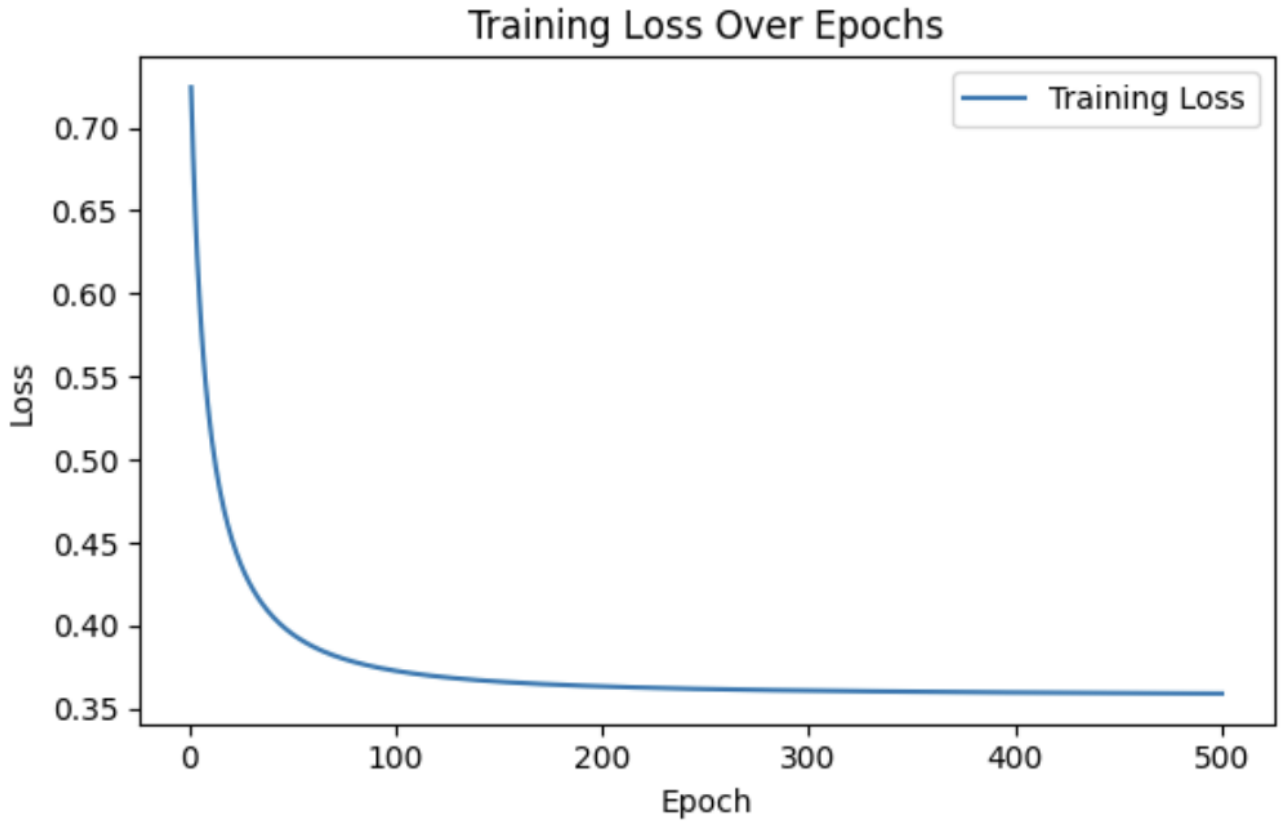
By running over 500 Epochs getting an accuracy of **71.90%**. This accuracy indicates the percentage of correctly predicted instances out of the total instances in our dataset.

The logistic regression model might not be performing as well as the deep neural network for a few reasons, logistic model has only one neuron to process the data, where in case for DNN to process we have N numbers of datapoints to process and generate the high accuracy.

Logistic regression assumes a linear relationship for binary classification, while deep neural networks excel in capturing complex patterns. Neural networks may overfit but capture intricate relationships. Logistic regression's simplicity resists overfitting but struggles with non-linear patterns, leading to lower accuracy. The choice depends on data complexity and the trade-off between model simplicity and predictive accuracy. From all these factors the linear regression model doesn't give us high accuracy.

```
Epoch [1/500], Training Loss: 0.7165, Training Accuracy: 42.59%
Epoch [2/500], Training Loss: 0.6783, Training Accuracy: 46.41%
Epoch [3/500], Training Loss: 0.6468, Training Accuracy: 53.94%
Epoch [4/500], Training Loss: 0.6206, Training Accuracy: 58.05%
Epoch [5/500], Training Loss: 0.5986, Training Accuracy: 60.59%
Epoch [6/500], Training Loss: 0.5798, Training Accuracy: 61.82%
Epoch [7/500], Training Loss: 0.5636, Training Accuracy: 63.53%
Epoch [8/500], Training Loss: 0.5495, Training Accuracy: 64.51%
Epoch [9/500], Training Loss: 0.5370, Training Accuracy: 65.32%
Epoch [10/500], Training Loss: 0.5260, Training Accuracy: 65.82%
...
Epoch [492/500], Training Loss: 0.3592, Training Accuracy: 71.75%
Epoch [493/500], Training Loss: 0.3592, Training Accuracy: 71.76%
Epoch [494/500], Training Loss: 0.3592, Training Accuracy: 71.76%
Epoch [495/500], Training Loss: 0.3592, Training Accuracy: 71.76%
Epoch [496/500], Training Loss: 0.3592, Training Accuracy: 71.76%
Epoch [497/500], Training Loss: 0.3591, Training Accuracy: 71.76%
Epoch [498/500], Training Loss: 0.3591, Training Accuracy: 71.76%
Epoch [499/500], Training Loss: 0.3591, Training Accuracy: 71.76%
Epoch [500/500], Training Loss: 0.3591, Training Accuracy: 71.76%
Accuracy on the test dataset: 0.7190165519714355
```


Training Accuracy Over Epochs
Accuracy
Epoch
Training Accuracy
0.70
0.65
0.60
0.55
0.50
0.45
0
100
200
300
400
500


Training Loss Over Epochs
Loss
Epoch
Training Loss
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0
100
200
300
400
500

d) Deep Neural Network in PyTorch

To train and test our DNN model, we'll utilize the feature vectors specified in TASK 1, specifically focusing on ***'num_dots'*** and '***url_len***'. The accuracy we are getting is above 95% which classifies as the best model to train and test on the Dataset.

In conclusion, while deep neural networks excel in capturing complex patterns and achieving high accuracy, logistic regression remains a viable choice for simpler problems where interpretability is paramount. The decision between these models should be driven by the specific task requirements, considering factors like dataset size, overfitting risks, and the need for straightforward interpretability. Ultimately, achieving high accuracy depends on a thoughtful selection of the appropriate model for the given scenario.

	url	type	num_dots	is_bad_url	url_len
0	br-icloud.com.br	phishing	2	False	16
1	mp3raid.com/music/krizz_kaliko.html	benign	2	False	35
2	bopsecrets.org/rexroth/cr/1.htm	benign	2	False	31
3	http://www.garage-pirenne.be/index.php?option=...	defacement	3	False	88
4	http://adventure-nicaragua.net/index.php?optio...	defacement	2	False	235
...	...	...	...	...	...
495	people.famouswhy.com/ramzi_yousef/	benign	2	False	34
496	wn.com/Raymond	benign	1	False	14
497	trtsport.cz	malware	1	False	11
498	youtube.com/watch?v=aYRauv5oeXQ	benign	1	False	31
499	armchairgm.wikia.com/San_Diego_Chargers	benign	2	False	39

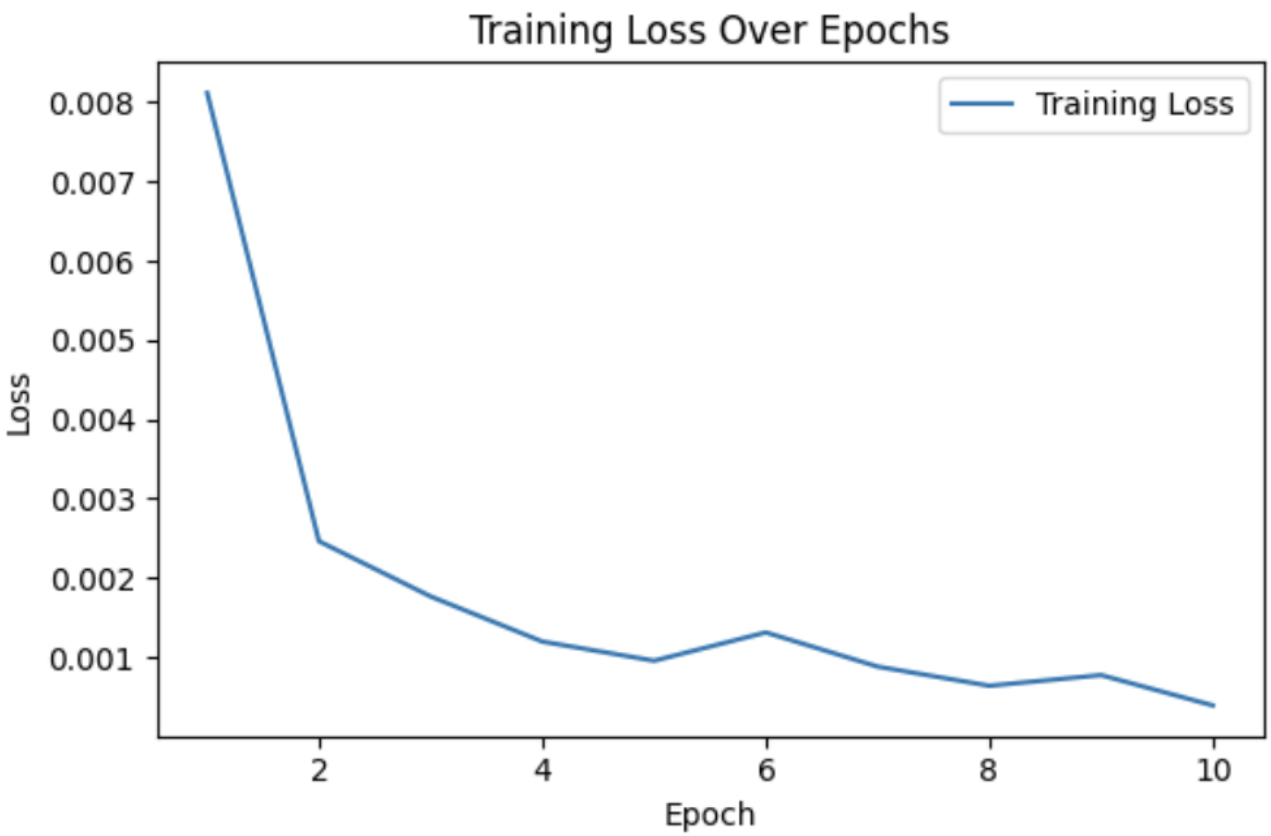
500 rows × 5 columns

	url	type	num_dots	is_bad_url	url_len
0	br-icloud.com.br	phishing	2	False	16
1	mp3raid.com/music/krizz_kaliko.html	benign	2	False	35
2	bopsecrets.org/rexroth/cr/1.htm	benign	2	False	31
3	http://www.garage-pirenne.be/index.php?option=...	defacement	3	False	88
4	http://adventure-nicaragua.net/index.php?optio...	defacement	2	True	235
...	...	...	...	...	...
495	people.famouswhy.com/ramzi_yousef/	benign	2	False	34
496	wn.com/Raymond	benign	1	False	14
497	trtsport.cz	malware	1	False	11
498	youtube.com/watch?v=aYRauv5oeXQ	benign	1	False	31
499	armchairgm.wikia.com/San_Diego_Chargers	benign	2	False	39

500 rows × 5 columns

```
Epoch [1/10], Training Loss: 0.0081, Training Accuracy: 99.80%
Epoch [2/10], Training Loss: 0.0025, Training Accuracy: 99.94%
Epoch [3/10], Training Loss: 0.0018, Training Accuracy: 99.95%
Epoch [4/10], Training Loss: 0.0012, Training Accuracy: 99.97%
Epoch [5/10], Training Loss: 0.0010, Training Accuracy: 99.98%
Epoch [6/10], Training Loss: 0.0013, Training Accuracy: 99.97%
Epoch [7/10], Training Loss: 0.0009, Training Accuracy: 99.98%
Epoch [8/10], Training Loss: 0.0006, Training Accuracy: 99.98%
Epoch [9/10], Training Loss: 0.0008, Training Accuracy: 99.99%
Epoch [10/10], Training Loss: 0.0004, Training Accuracy: 99.99%
Testing Accuracy: 100.00%
```

> Plotting the results.


Training Loss Over Epochs
Training Loss
Loss
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
2
4
6
8
10
Epoch