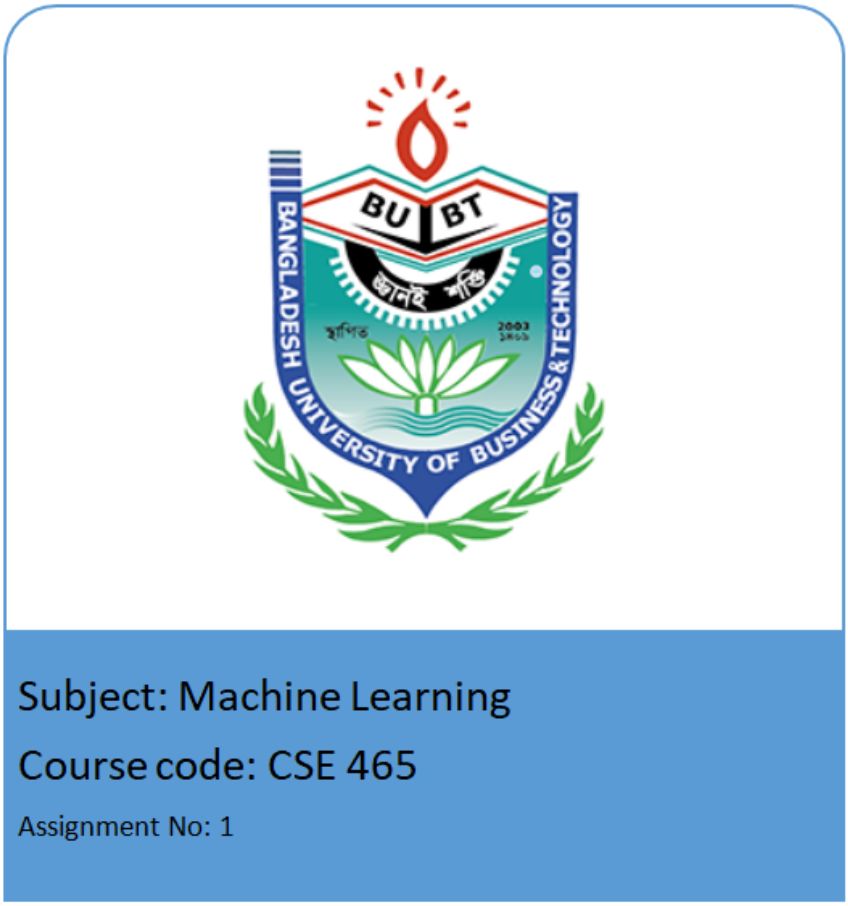
Assignment 1



Submitted To	Submitted By
Khan Md. Hasib Lecturer Dept. of CSE BUBT	Ahmed Mahir Shoaib ID: 18192103235 Intake: 41 Section: 7 Dept. of CSE BUBT

> Apply data preprocessing steps (such as: Viewing your data, Handling
> duplicates, Column cleanup, DataFrame slicing, selecting, extracting) in
> the following dataset -
> https://www.kaggle.com/datasets/selinraja/irish-
> data .

Here are the steps for data preprocessing.

1. Viewing your data:

We can start by loading the dataset into a pandas DataFrame and taking a look at the first few rows using the head() method:

```
import pandas as pd

df = pd.read_csv('/content/archive.zip')
print(df.head())
```

This will print the first five rows of the dataset to the console.

2. Handling duplicates:

To check for duplicates in the dataset, we can use the duplicated() method:

```
duplicates = df.duplicated()
print(duplicates.sum())
```

This will print the number of duplicate rows in the dataset. If we want to remove the duplicates, we can use the drop_duplicates() method:

```
df = df.drop_duplicates()
```

This will remove all duplicate rows from the DataFrame.

3. Column cleanup:
   The column names in the dataset already look clean, so we don't need to

perform any cleanup.

4. DataFrame slicing, selecting, extracting:
   To slice the DataFrame and select specific rows and columns, we can use the loc[] method:

```
subset = df.loc[10:20, ['sepal_length', 'petal_length', 'species']]
```

This will select rows 10 to 20 and columns 'sepal_length', 'petal_length', and 'species' from the DataFrame.

To select rows based on a condition, we can use boolean indexing:

```
versicolor = df[df['species'] == 'versicolor']
print(versicolor)
```

This are the process for data preprocessing.