Assignment 6 – Model Training

DS 561 – Cloud Computing

Vishwas Bhaktavatsala
U74206902
vishwasb@bu.edu
GitHub Url: https://github.com/vishwas21/DS561-vishwas-assignments

---

Model used to predict the country is RandomForestClassifier and below is the accuracy which I achieved.

```
[vishwasb@assignment6-model-vm:~$ python3 countryModel.py
 Model accuracy: 100.00%
 vishwasb@assignment6-model-vm:~$
```

Model used to predict the income is RandomForestClassifier again and below is the accuracy which I achieved.

```
[vishwasb@assignment6-model-vm:~$ python3 incomeModel.py
 Model accuracy: 45.12%
 Model accuracy: 45.12%
```

New Accuracy:

```
vishwasb@assignment6-model-vm:~$ python3 incomeModel.py
Model accuracy: 45.79%
Model accuracy: 45.79%
vishwasb@assignment6-model-vm:~$
```

Note: I was trying to get a better accuracy, but unfortunately forgot to upload code and this document. Hope this is not considered as a late submission, because I had already submitted on gradescope before the due date.

Below are the steps to follow:

Download the code from the Git Repository using the Git Clone Command

cd into the assignment6 folder

Run the command pip3 install -r requirements.txt

Once the installation is complete, run the following commands to run each model

python3 countryModel.py

python3 incomeModel.py

RandomForestClassifier:
RandomForestClassifier is an ensemble machine learning algorithm used for classification tasks. It builds multiple decision trees during training and combines their predictions through a majority vote (for classification) to improve accuracy and reduce overfitting. It also introduces randomness by using random subsets of features and data, making it robust and versatile. It's a popular choice for various classification problems due to its effectiveness and ability to handle large datasets.

Country Model:
My Code is trying to find the correlations between the country and the ip address with 80% of the dataset and train the model. Using this model I tested with the rest of the data and calculate the Accuracy.

Income Model:
My code first would change the range values of Age and Income to numerical values and Label encode the country so that we can work completely with numerical values.

Using these values, I am training my model with the values from the columns ip, gender and age using which I would be trying to calculate the income. Once the model is trained, I will use it to test with the rest of the values and predict the income from my test set. Initially, I got a lesser accuracy, but now I could achieve an accuracy of 45.79%.