

DS561-HW6

U43188754

Github Link:

<https://github.com/tigeryi1998/ds561-tigeryi/tree/main/hw6>

0. SQL Database “dbhw5” from HW5

<https://console.cloud.google.com/sql/instances/instance-tigeryi/overview?project=feisty-gasket-398719>

Public IP address

34.138.218.160

```
DB_NAME = "dbhw5"
INSTANCE_CONNECTION_NAME = "feisty-gasket-398719:us-east1:instance-tigeryi"
```

table1 (successful requests)

```
mysql> select * from table1 limit 5;
+-----+-----+-----+-----+-----+
| request_id | ip          | time_of_day      | filename | ip2          |
+-----+-----+-----+-----+-----+
| 1          | 7.176.22.73 | 2023-11-08 14:00:00 | 6783.html | 128980553    |
| 2          | 9.45.75.21  | 2023-11-08 21:00:00 | 2872.html | 153963285    |
| 3          | 172.14.168.211 | 2023-11-08 08:00:00 | 5130.html | 2886641875   |
| 4          | 67.184.165.29 | 2023-11-08 12:00:00 | 8503.html | 1136174365   |
| 5          | 206.117.156.36 | 2023-11-08 06:00:00 | 4090.html | 3463814180   |
+-----+-----+-----+-----+-----+
5 rows in set (0.01 sec)
```

table2

```
mysql> select * from table2 limit 2;
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| ip_id | ip          | gender | age  | income | country | is_banned | ip2          | gender2 | age2 | income2 | country2 |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 1      | 7.176.22.73 | Male   | 36-45 | 20k-40k | Nicaragua | 0         | 128980553    | 0       | 3    | 2        | 124      |
| 2      | 9.45.75.21  | Male   | 36-45 | 100k-150k | Cote d'Ivoire | 0         | 153963285    | 0       | 3    | 5        | 41       |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
2 rows in set (0.00 sec)
```

table3 (failed requests)

```
mysql> select * from table3 limit 5;
+-----+-----+-----+-----+-----+-----+
| failed_id | ip          | time_of_day      | filename | error | ip2          |
+-----+-----+-----+-----+-----+-----+
| 1          | 172.237.141.166 | 2023-11-08 08:00:00 | 12257.html | 404 | 2901249446   |
| 2          | 165.7.115.176  | 2023-11-08 17:00:00 | 10123.html | 404 | 2768729008   |
| 3          | 165.172.187.193 | 2023-11-08 14:00:00 | 19946.html | 404 | 2779560897   |
| 4          | 115.215.109.81  | 2023-11-08 14:00:00 | 18713.html | 404 | 1943498065   |
| 5          | 18.128.133.193  | 2023-11-08 00:00:00 | 11589.html | 404 | 310412737    |
+-----+-----+-----+-----+-----+-----+
5 rows in set (0.00 sec)
```

Database Connection with Pandas Dataframe

Details in `app1.ipynb` and `app2.ipynb`

```
PROJECT_ID = "feisty-gasket-398719"
TOPIC_ID = "my-topic"
SUBSCRIPTION_NAME = "my-topic-sub"
INSTANCE_CONNECTION_NAME = "feisty-gasket-398719:us-east1:instance-tigeryi"
DB_USER = "root"
DB_PASS = ""
DB_NAME = "dbhw5"
DB_PRIVATE_IP = False

def connect_with_connector() -> sqlalchemy.engine.base.Engine:
...
```

I will use Pandas with SQL Alchemy to read the database table into Pandas dataframe

1. model1 (ip, country)

Code for the model is in Jupyter Notebook
`app1.ipynb`

```
pool = connect_with_connector()
query = '''select ip2, country2, ip, country from table2;'''
df = pd.read_sql_query(query, pool)
print(df.head())
```

	ip2	country2	ip	country
0	2737397699	160	163.41.95.195	Somalia
1	2509268020	89	149.144.100.52	Kosovo
2	1314065103	113	78.83.10.207	Monaco
3	2436094690	9	145.51.218.226	Austria
4	294390865	162	17.140.12.81	South Korea

Model is decision tree classifier `DecisionTreeClassifier`

```
X = df[['ip2']]
y = df['country2']
```

```
clf = DecisionTreeClassifier()
clf.fit(X,y)
```

```
clf.score(X,y)
1.0
y_pred = clf.predict(X)
accuracy_score(y, y_pred)
1.0
```

The decision tree model has 100% accuracy rate, which is really good

2. model2 (income, gender, age, country)

Code for the model is in Jupyter Notebook
app2.ipynb

```
pool = connect_with_connector()
query = '''select ip2, country2, gender2, age2, income2, ip, country, gender, age,
income from table2;'''
df = pd.read_sql_query(query, pool)
print(df.head())
```

	ip2	country2	gender2	age2	income2	ip	country	\
0	2737397699	160	1	0	6	163.41.95.195	Somalia	
1	2509268020	89	1	0	6	149.144.100.52	Kosovo	
2	1314065103	113	0	1	1	78.83.10.207	Monaco	
3	2436094690	9	0	7	0	145.51.218.226	Austria	
4	294390865	162	1	1	4	17.140.12.81	South Korea	

	gender	age	income
0	Female	0-16	150k-250k
1	Female	0-16	150k-250k
2	Male	17-25	10k-20k
3	Male	76+	0-10k
4	Female	17-25	60k-100k

Model is decision tree classifier `RandomForestClassifier`

```
X = df[['ip2', 'country2', 'gender2', 'age2']]
y = df['income2']
```

```
clf = RandomForestClassifier(n_estimators=200)
clf.fit(X,y)
clf.score(X,y)
0.7450418649068058
```

```
y_pred = clf.predict(X)
accuracy_score(y, y_pred)
0.7450418649068058
```

The random forest model has 74.5% accuracy rate, close to the 80% threshold