## **BUILDING THE TRAINED MODEL**

DATA PREPROCESSING

```
In [1]:
         import pandas as pd
In [2]:
         dataset = pd.read csv('../dataset/train dataset.csv', index col = 0)
In [3]:
         dataset.head()
Out[3]:
           trans_hour trans_day trans_month trans_year category card_number age trans_amount state
                                                                                                     zip fraud_risk
        0
                   0
                                                2019
                                                           12 6.300000e+11
                                                                                       66.21
                                                                                               22 49879
                                                                                                                0
                                                2019
                                                            3 3.540000e+15
                                                                                       55.81
                                                                                               14 62668
                                                2019
                                                            8 5.020000e+11
                                                                                        8.68
                                                                                                4 96037
                            1
                                        1
                                                2019
                                                            4 3.530000e+15
                                                                                       89.52
                                                                                               40 29911
                   6
                            1
                                        1
                                                2019
                                                            0 2.350000e+15
                                                                           72
                                                                                        1.90
                                                                                               38 16421
                                                                                                                0
In [4]:
         import numpy as np
In [5]:
         x = dataset.iloc[ : , : 10].values
         y = dataset.iloc[ : , 10].values
In [6]:
         from sklearn.model selection import train test split
In [7]:
         x train, x test, y train, y test = train test split(x, y, test size = 0.15, random state = 0)
In [8]:
         x train.shape
```

```
Out[8]: (10210, 10)
 In [9]:
          x test.shape
Out[9]: (1802, 10)
In [10]:
          fraud = np.count nonzero(y train == 1)
          valid = np.count nonzero(y train == 0)
In [11]:
          print('Fraud cases in training data =', fraud)
          print('Valid cases in training data =', valid)
         Fraud cases in training data = 5106
         Valid cases in training data = 5104
In [12]:
          from sklearn.preprocessing import StandardScaler
          scaler = StandardScaler()
In [13]:
          x train = scaler.fit transform(x train)
          x test = scaler.transform(x test)
In [14]:
          x train[:5]
Out[14]: array([[ 0.66337259, 1.24104545, 0.45653094, -0.37826057, 0.83408097,
                 -0.30554032, 2.3966382, -0.73068363, 0.0181697, -0.77660466],
                [-1.6023747 , -0.07304053 , -0.90455674 , -0.37826057 , -1.22335066 ,
                 -0.30836573, 0.74498916, -0.74861649, 0.64443135, 0.93890762],
                [-1.12537527, 0.1459738, 0.45653094, -0.37826057, 0.57690202,
                 -0.30362247, 1.18542891, -0.79511806, 0.78360061, -1.2179066 ],
                [-1.6023747, -0.29205486, -0.36012167, -0.37826057, -0.70899275,
                 -0.30397763, 1.46070375, -0.51118115, -1.58227676, 1.76478537],
                [-0.7676257, -0.29205486, -1.44899181, -0.37826057, -0.70899275,
                 -0.30836532, -0.08083536, 0.02559291, 0.78360061, -1.18526965]])
In [15]:
          x test[ : 5]
Out[15]: array([[ 1.02112216, -0.51106919, 1.27318354, -0.37826057, 1.34843888,
                  3.15638262, 0.35960438, 1.60624238, 0.64443135, 0.91726066],
```

```
[-0.05212655, 0.58400246, 0.1843134, -0.37826057, 0.83408097,
                 -0.30822372, -0.08083536, -0.65192446, 0.29650821, -1.52769864],
                 [ 1.14037202, 1.56956694, 0.72874847, -0.37826057, 1.09125992,
                 -0.30560345, 1.29553884, 1.71351642, 1.13152375, 1.07356281],
                [1.02112216, -0.73008352, 0.1843134, -0.37826057, 1.09125992,
                 -0.30500363, 1.51575871, 1.6205402, 0.15733896, 0.77172649],
                 [-1.6023747 , -0.07304053 , -0.08790413 , -0.37826057 , 1.34843888 ,
                 -0.30819767, 0.80004413, -0.06571388, -0.60809196, 0.84854544]])
         LOGISTIC REGRESSION (LR)
In [16]:
          from sklearn.linear model import LogisticRegression
          LR model = LogisticRegression(random state = 0)
          LR model.fit(x train, y train)
Out[16]: LogisticRegression(random_state=0)
In [17]:
          y pred = LR model.predict(x test)
In [18]:
          from sklearn.metrics import accuracy score
In [19]:
          acc lr = accuracy score(y test, y pred)
In [20]:
          print(acc lr)
         0.8485016648168702
         K-NEAREST NEIGHBORS (KNN)
In [21]:
          from sklearn.neighbors import KNeighborsClassifier
          KNN model = KNeighborsClassifier(n neighbors = 5, metric = 'minkowski', p = 2)
          KNN model.fit(x train, y train)
Out[21]: KNeighborsClassifier()
In [22]:
          y pred = KNN model.predict(x test)
```

```
acc_knn = accuracy_score(y_test, y_pred)
In [23]:
In [24]:
          print(acc knn)
         0.8657047724750278
         SUPPORT VECTOR MACHINE (SVM)
In [25]:
          from sklearn.svm import SVC
          SVM_model = SVC(kernel = 'linear', random_state = 0)
          SVM_model.fit(x_train, y_train)
Out[25]: SVC(kernel='linear', random_state=0)
In [26]:
          y pred = SVM model.predict(x test)
In [27]:
          acc_svm = accuracy_score(y_test, y_pred)
In [28]:
          print(acc svm)
          0.8479467258601554
         NAIVE BAYES (NB)
In [29]:
          from sklearn.naive_bayes import GaussianNB
          NB model = GaussianNB()
          NB_model.fit(x_train, y_train)
Out[29]: GaussianNB()
In [30]:
          y pred = NB model.predict(x test)
In [31]:
          acc_nb = accuracy_score(y_test, y_pred)
In [32]:
          print(acc_nb)
```

```
DECISION TREE (DT)
In [33]:
          from sklearn.tree import DecisionTreeClassifier
          DT_model = DecisionTreeClassifier(criterion = 'entropy', random_state = 0)
          DT_model.fit(x_train, y_train)
Out[33]: DecisionTreeClassifier(criterion='entropy', random_state=0)
In [34]:
          y pred = DT model.predict(x test)
In [35]:
          acc_dt = accuracy_score(y_test, y_pred)
In [36]:
          print(acc_dt)
          0.967258601553829
         RANDOM FOREST (RF)
In [37]:
          from sklearn.ensemble import RandomForestClassifier
          RF model = RandomForestClassifier()
          RF_model.fit(x_train, y_train)
Out[37]: RandomForestClassifier()
In [38]:
          y pred = RF model.predict(x test)
In [39]:
          acc_rf = accuracy_score(y_test, y_pred)
In [40]:
          print(acc rf)
          0.9705882352941176
         ARTIFICIAL NEURAL NETWORK (ANN)
```

0.8496115427302997

```
import tensorflow as tf
In [41]:
In [42]:
   ANN model = tf.keras.models.Sequential()
In [43]:
   ANN model.add(tf.keras.layers.Dense(64, input dim = 10, activation = 'relu'))
   ANN model.add(tf.keras.layers.Dense(128, activation = 'relu'))
   ANN model.add(tf.keras.layers.Dense(1, activation = 'sigmoid'))
In [44]:
   ANN model.compile(optimizer = 'adam', loss = 'binary crossentropy', metrics = ['accuracy'])
In [45]:
   ANN model.fit(x train, y train, batch size = 32, epochs = 200)
   Epoch 1/200
   Epoch 2/200
   Epoch 3/200
   Epoch 4/200
   Epoch 5/200
   Epoch 6/200
   Epoch 7/200
   Epoch 8/200
   Epoch 9/200
   Epoch 10/200
   Epoch 11/200
   Epoch 12/200
   Epoch 13/200
   Epoch 14/200
   Epoch 15/200
```

```
Epoch 16/200
v: 0.9156
Epoch 17/200
Epoch 18/200
Epoch 19/200
Epoch 20/200
Epoch 21/200
Epoch 22/200
Epoch 23/200
Epoch 24/200
Epoch 25/200
Epoch 26/200
Epoch 27/200
Epoch 28/200
Epoch 29/200
Epoch 30/200
Epoch 31/200
Epoch 32/200
Epoch 33/200
Epoch 34/200
Epoch 35/200
Epoch 36/200
Epoch 37/200
Epoch 38/200
Epoch 39/200
```

```
Epoch 40/200
Epoch 41/200
Epoch 42/200
Epoch 43/200
Epoch 44/200
Epoch 45/200
Epoch 46/200
Epoch 47/200
Epoch 48/200
Epoch 49/200
Epoch 50/200
Epoch 51/200
Epoch 52/200
Epoch 53/200
Epoch 54/200
Epoch 55/200
Epoch 56/200
Epoch 57/200
Epoch 58/200
Epoch 59/200
Epoch 60/200
Epoch 61/200
Epoch 62/200
Epoch 63/200
Epoch 64/200
```

```
Epoch 65/200
Epoch 66/200
Epoch 67/200
Epoch 68/200
Epoch 69/200
Epoch 70/200
Epoch 71/200
Epoch 72/200
Epoch 73/200
Epoch 74/200
Epoch 75/200
Epoch 76/200
Epoch 77/200
Epoch 78/200
Epoch 79/200
Epoch 80/200
Epoch 81/200
Epoch 82/200
Epoch 83/200
Epoch 84/200
Epoch 85/200
Epoch 86/200
Epoch 87/200
Epoch 88/200
```

```
Epoch 89/200
Epoch 90/200
Epoch 91/200
Epoch 92/200
Epoch 93/200
Epoch 94/200
Epoch 95/200
Epoch 96/200
Epoch 97/200
Epoch 98/200
Epoch 99/200
Epoch 100/200
Epoch 101/200
Epoch 102/200
Epoch 103/200
Epoch 104/200
Epoch 105/200
Epoch 106/200
Epoch 107/200
Epoch 108/200
Epoch 109/200
Epoch 110/200
Epoch 111/200
Epoch 112/200
Epoch 113/200
```

```
Epoch 114/200
Epoch 115/200
Epoch 116/200
Epoch 117/200
Epoch 118/200
Epoch 119/200
Epoch 120/200
Epoch 121/200
Epoch 122/200
Epoch 123/200
Epoch 124/200
Epoch 125/200
Epoch 126/200
Epoch 127/200
Epoch 128/200
Epoch 129/200
Epoch 130/200
Epoch 131/200
Epoch 132/200
Epoch 133/200
Epoch 134/200
Epoch 135/200
Epoch 136/200
Epoch 137/200
```

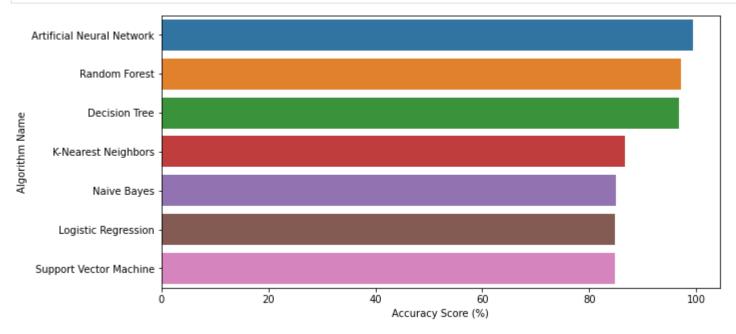
```
Epoch 138/200
Epoch 139/200
Epoch 140/200
Epoch 141/200
Epoch 142/200
Epoch 143/200
Epoch 144/200
Epoch 145/200
Epoch 146/200
Epoch 147/200
Epoch 148/200
Epoch 149/200
Epoch 150/200
Epoch 151/200
Epoch 152/200
Epoch 153/200
Epoch 154/200
Epoch 155/200
v: 0.9851
Epoch 156/200
Epoch 157/200
Epoch 158/200
Epoch 159/200
Epoch 160/200
Epoch 161/200
```

```
Epoch 162/200
Epoch 163/200
Epoch 164/200
Epoch 165/200
Epoch 166/200
Epoch 167/200
Epoch 168/200
Epoch 169/200
Epoch 170/200
Epoch 171/200
Epoch 172/200
Epoch 173/200
Epoch 174/200
Epoch 175/200
Epoch 176/200
Epoch 177/200
Epoch 178/200
Epoch 179/200
Epoch 180/200
Epoch 181/200
Epoch 182/200
Epoch 183/200
Epoch 184/200
Epoch 185/200
Epoch 186/200
```

```
Epoch 187/200
  Epoch 188/200
  Epoch 189/200
  Epoch 190/200
  Epoch 191/200
  Epoch 192/200
  Epoch 193/200
  Epoch 194/200
  Epoch 195/200
  Epoch 196/200
  Epoch 197/200
  Epoch 198/200
  Epoch 199/200
  Epoch 200/200
  Out[45]: <tensorflow.python.keras.callbacks.History at 0x1c56862beb0>
In [46]:
   loss, acc ann = ANN model.evaluate(x train, y train, verbose = 0)
In [47]:
   print(acc ann)
  0.994025468826294
In [48]:
  v pred = ANN model.predict(x test)
  y pred[y pred <= 0.5] = 0</pre>
  y pred[y pred > 0.5] = 1
  ACCURACY COMPARISON OF ALL THE MODELS
```

```
scores = [acc_lr * 100,
                      acc_knn * 100,
                      acc_svm * 100,
                      acc_nb * 100,
                      acc_dt * 100,
                      acc_rf * 100,
                      acc ann * 100]
In [50]:
           names = ["Logistic Regression",
                    "K-Nearest Neighbors",
                    "Support Vector Machine",
                    "Naive Bayes",
                    "Decision Tree",
                    "Random Forest",
                    "Artificial Neural Network"]
In [51]:
           df = pd.DataFrame()
           df['Algorithm Name'] = names
           df['Accuracy Score (%)'] = scores
           df = df.sort_values('Accuracy Score (%)', ascending = False)
In [52]:
           df
Out[52]:
                   Algorithm Name Accuracy Score (%)
          6 Artificial Neural Network
                                          99.402547
          5
                     Random Forest
                                          97.058824
                                          96.725860
                      Decision Tree
                K-Nearest Neighbors
                                          86.570477
          3
                       Naive Bayes
                                          84.961154
          0
                  Logistic Regression
                                          84.850166
          2 Support Vector Machine
                                          84.794673
In [53]:
           import matplotlib.pyplot as plt
           import seaborn as sns
```

```
fig = plt.subplots(figsize = (10, 5))
ax = sns.barplot(x = "Accuracy Score (%)", y = "Algorithm Name", data = df)
```



## SAVING THE BEST TRAINED MODEL