Stats 101C Homework 6

Damien Ha

In [1]:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
```

(1)

Randomly Sample 10,000 data points as training data and let the rest of the data be the testing dataset. Build a decision tree (Maximum Depth = 5) on 10,000 data points and test its performance on the testing data.

In [2]:

```
df = pd.read_csv('Smoke_Alarm_Dataset.csv')
X = df.drop('Fire Alarm', axis=1)
y = df['Fire Alarm']
```

In [3]:

```
# Randomly sample 10,000 data points for training
X_train, X_test, y_train, y_test = train_test_split(X, y, train_size = 10000, random_state=1)

# Build Decision Tree model with maximum depth = 5
decision_tree_model = DecisionTreeClassifier(max_depth=5)
decision_tree_model.fit(X_train, y_train)

# Evaluate Decision Tree model on testing data
y_pred_dt = decision_tree_model.predict(X_test)
accuracy_dt = accuracy_score(y_test, y_pred_dt)
print(f"Decision Tree Accuracy: {accuracy_dt:.4f}")
```

```
Decision Tree Accuracy: 0.9861
```

(2)

Build a random forest model (number of trees = 15) on the training dataset and test its performance on the testing data.

In [4]:

```
# Build Random Forest model with 15 trees
random_forest_model = RandomForestClassifier(n_estimators=15, random_state=1)
random_forest_model.fit(X_train, y_train)

# Evaluate Random Forest model on testing data
y_pred_rf = random_forest_model.predict(X_test)
accuracy_rf = accuracy_score(y_test, y_pred_rf)
print(f"Random Forest Accuracy: {accuracy_rf:.4f}")
```

```
Random Forest Accuracy: 0.9997
```

(3)

Compare the accuracy of decision trees and random forests. What is your conclusion?

In [5]:

```
print("\nComparison:")
print(f"Decision Tree Accuracy: {accuracy_dt:.4f}")
print(f"Random Forest Accuracy: {accuracy_rf:.4f}")

# Conclusion
if accuracy_rf > accuracy_dt:
    print("Random Forest outperforms Decision Tree.")
elif accuracy_rf < accuracy_dt:
    print("Decision Tree outperforms Random Forest.")
else:
    print("Decision Tree and Random Forest have similar performance.")
```

```
Comparison:
Decision Tree Accuracy: 0.9861
Random Forest Accuracy: 0.9997
Random Forest outperforms Decision Tree.
```