Report

1. Description

The dataset called “occupancy” downloaded from Kaggle dataset for this particular project.

Firstly, we imported all the necessary libraries and classes.

The dataset originally included 8 columns called: id, date, Temperature, Humidity, Light, CO2, HumidityRatio, and Occupancy. The dataset had total 17,895 rows. I changed the date column into 5 new columns: year, month, day, hour, and minute and dropped original date column. Then I had total 12 columns. I used klib data cleaning formula to change standardize column names, drop duplicated rows and reduced memory usage from 1.1MB to 0.56MB (-51.38%). In the data analysis part, I marked occupancy as our target value (y) and remaining 11 columns as input values (X).

This dataset belongs to classification model as the target value has only 2 unique values: 0 and 1.

I found the count values of 1 and 0 in the target column:

```
Occupancy
0       14117
1        3778
Name: count, dtype: int64
```

2. Model

We applied random forest classifier model for this dataset.

We marked target value: column called occupancy as y and all the remaining columns as X.

We used standard scaler to normalize input values.

We split the given dataset into 2 parts: test 30% and train 70%.

After training the model, we found accuracy with accuracy score.

Our accuracy score was: 0.9955307262569832

3. Hyperparameter tuning

To improve the model, we used hyper parameter tuning method.

Firstly, we assigned our parameters as below:

```
param_grid = {
    'n_estimators': [60, 120, 300],
    'max_depth': [6, 12, 30],
    'min_samples_split': [2, 5, 10]
}
```

Then we used GridSearchCV to find best parameters for our random forest model.

And the best parameters were as below:

{'max_depth': 30, 'min_samples_split': 2, 'n_estimators': 120}
Upon applying these parameters on random forest model, we increased the accuracy to 0.9962756052141527.

Next we used Random Search CV for the same purpose.

Our assigned parameters for Random Search CV were

```
'n_estimators': randint(50, 300),
'max_depth': randint(5, 30),
'min_samples_split': randint(2, 20),
'min_samples_leaf': randint(1, 10)
```

And we got best parameters as below: {'max_depth': 28, 'min_samples_leaf': 3, 'min_samples_split': 3, 'n_estimators': 137}

Then we applied these figures on random forest model and got the accuracy level of 0.9930148086057557