Determining space occupancy based on room variables

UHSE Machine learning boot camp 2020
(Tony Tran, Garima Singh, Erika Andrade, Mark Martir- Irizarry)

Problem Introduction

Determining the occupancy of a person in a room has many advantages in terms of learning the natural habits of an employee in an office, or even a family member in a home. This information will greatly benefit companies who are looking for a way to make the workplace and home more productive or want to save some money in terms of reducing energy cost.

The main goals of this study are:

(1) Predict room occupancy using the given variables
(2) Study the correlation of these factors with room occupancy and with themselves
(3) Find out which variables explain the Occupancy trends sufficiently? does light, CO2, humidity, temperature affect if someone is in the room or not, and how effective are those conditions?
(4) Evaluate the effectiveness and accuracy of the different machine learning techniques on this dataset

Dataset

The dataset used for this project was obtained from UC Irvine Machine Learning Repository. The data set consists of multiple samples of the room's condition, as well as the room occupancy during that time. With all the information, the shape of the dataset was a 2665 by 6 with no missing values.

Features and Processing

The list of features used to determine the occupancy of a room are as below:

1) **Temperature**, in Celsius
2) **Relative Humidity**, %
3) **Light**, in Lux
4) **CO2**, in ppm
5) **Humidity Ratio**, Derived from temperature and relative humidity, in kgwater-vapor/kg-air

We initially plotted the histogram of the dataset to analyze if the dataset was normally distributed and check to make sure that there were no missing datapoint or outliers that might cause future errors. After the confirmation no extreme outliers, errors, or blanks in the data, we then proceeded to plot the features while setting the color hue to match the occupancy. By doing so, we are able to visualize how the features

are related to each other, and also see if we can visually notice a distinction between the lines that determine if someone is residing in the room based on those criteria, as shown in figure 1. In many of the cases, since our dataset were plotted using two features at a time, there are a bunch of overlaps, which shows that two factors alone does not fully determine the room occupancy.

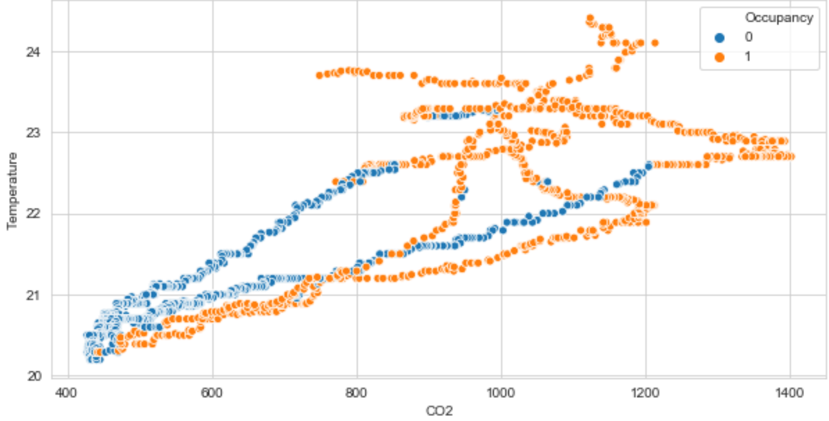
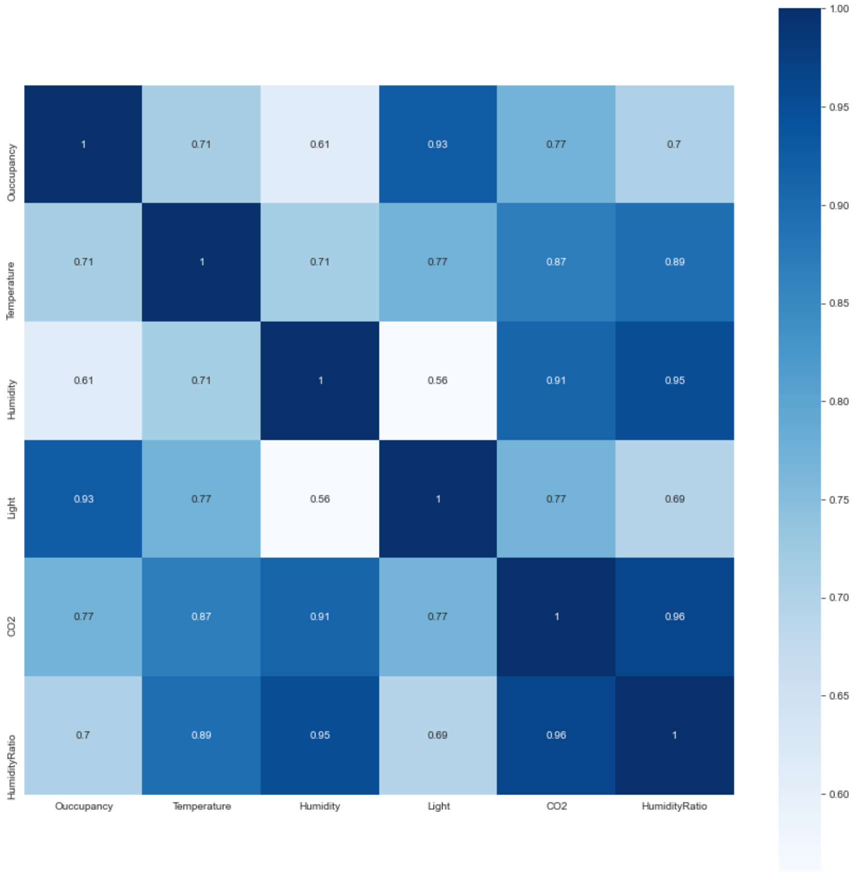


Figure 1: Scatterplot of CO2 V.S. Temperature (orange: room occupied, blue: room empty)

We then applied a correlation matrix, which was used to determine how much each of the features are related to each other. After some splitting and scaling of the dataset based on the training inputs, we were able to apply that data and create the heatmap shown in Figure 2. Based on the figure below, we should note that the 'Humidity ratio' has a high correlation with both the 'Humidity (corr: 0.95)' and the 'Concentration of CO2 (corr: 0.96)', as well as a close correlation with the 'Temperature (corr: 0.89)'. This correlation makes sense because the humidity ratio was determined using data from humidity and temperature dataset. It is also interesting to note that the 'light' has the least correlation with the other features in the dataset (average corr: 0.7), and yet has a highest correlation with the 'Room occupancy (corr: 0.93)'.



	Oucupancy	Temperature	Humidity	Light	CO2	HumidityRatio
Oucupancy	1	0.71	0.61	0.93	0.77	0.7
Temperature	0.71	1	0.71	0.77	0.87	0.89
Humidity	0.61	0.71	1	0.56	0.91	0.95
Light	0.93	0.77	0.56	1	0.77	0.69
CO2	0.77	0.87	0.91	0.77	1	0.96
HumidityRatio	0.7	0.89	0.95	0.69	0.96	1

Figure 2 Correlation heatmap of all the features

Models and techniques

We split the data set into training set and training set in the ratio 80: 20 and use a variety of Machine Learning techniques to predict the occupancy of a room. The list of machine learning that we used for this project is linear and nonlinear SVM, random forest, and Neural networks. We also used techniques such as pipeline and grid search to modify the parameters of each machine learning algorithm to see which parameter gives the most accurate results. By using a diverse set of techniques, we can determine the most optimal solution, as well as which of those algorithms fit best with the current data set and criteria.

Results and Discussion

The determination of a room occupancy was accurately measured using the given features throughout all the different Machine learning technique on average with an accuracy of over 95% as shown in Figure 3. This accuracy was obtained by using a confusion matrix which count the number of times the model accurately classify the room occupancy in a 2 by 2 matrix, as well as the built in scoring system which calculate the percentage of correct estimation with respect to the testing dataset. The confusion matrix has the format of [[nn,ny] [yn,yy]], where the first letter represent the actual yes or no and the second letter represent the predicted yes or no to the room occupancy. In other words, if we need to determine how many times the model predicted that there was someone in the room(‘y’) when there was actually no one(‘n’), we would look at the number in the ‘ny’ position. This information is useful in calculating the accuracy score, as well as finding out what causes most of the errors in the specific trained model.

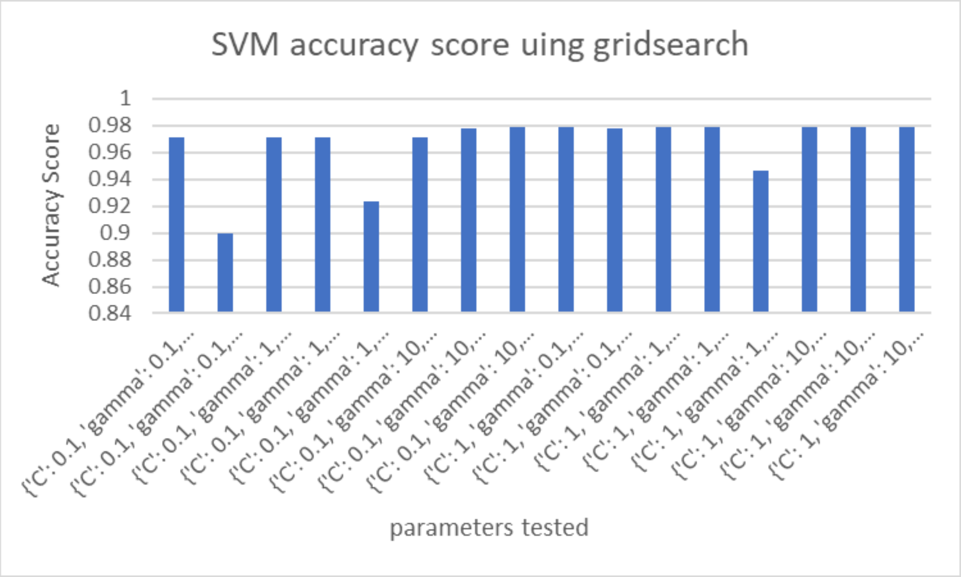


Figure 3: Example of the parameters tested using SVM grid search and the accuracy score for each test

The predictions of room occupancy using the best fitness model of each of the different machine learning approaches and parameters are shown in Figure -6. Based on the figures below, it is clear that each method uses a different shape to classify room occupancy with different success rates. It is important to note that while it looks like each method’s fitness model does not accurately show the true occupancy in a 2d plot, the accuracy is over 94% for each case. In higher dimensions (using all five features) it is 96-98%. Therefore, PCA can provide computational advantage and better visualisation when the data set is large with many features, if a 3% drop in prediction accuracy is acceptable.

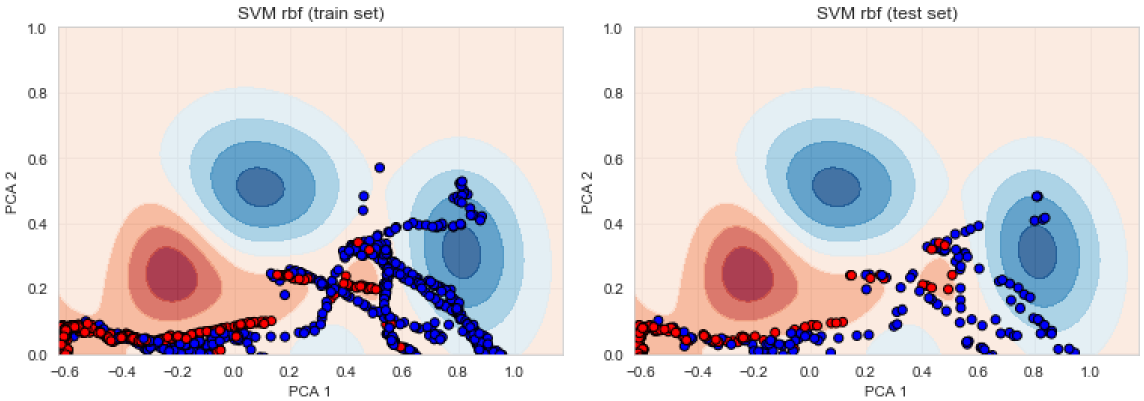


Figure 4: Prediction classification of room occupancy based on Neural Networks (Blue is not occupied, Red is occupied). The accuracy score for this fitness model is: 0.979

For SVM, the grid search determined the most optimal solution with {'C': 1000, 'gamma': 10, 'kernel': 'rbf'}, with a confusion matrix of [[338,11][0,184]]. This means that by using the specified parameters for this model, it will give a high accuracy score, and all the errors in this model comes from misclassifying occupancy when there was no one in the room.

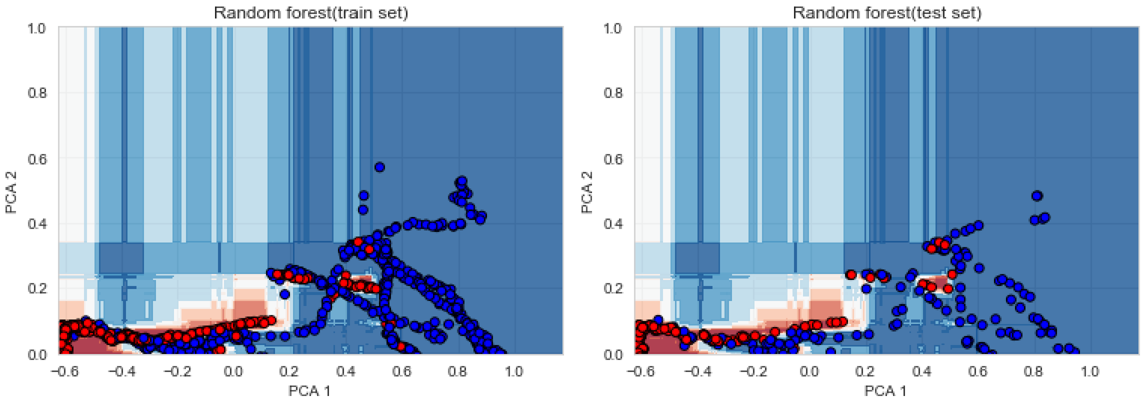


Figure 5: Prediction classification of room occupancy based on Neural Networks (Blue is not occupied, Red is occupied). The accuracy score for this fitness model is: 0.983

For Random Forest, the grid search determined the most optimal solution with {'max depth': 8, 'K': 2}, with a confusion matrix of [[341,8][1,183]]. This means that by using the specified parameters for this model, it will give a high accuracy score, and all the errors in this model mostly comes from misclassifying occupancy when there was no one in the room.

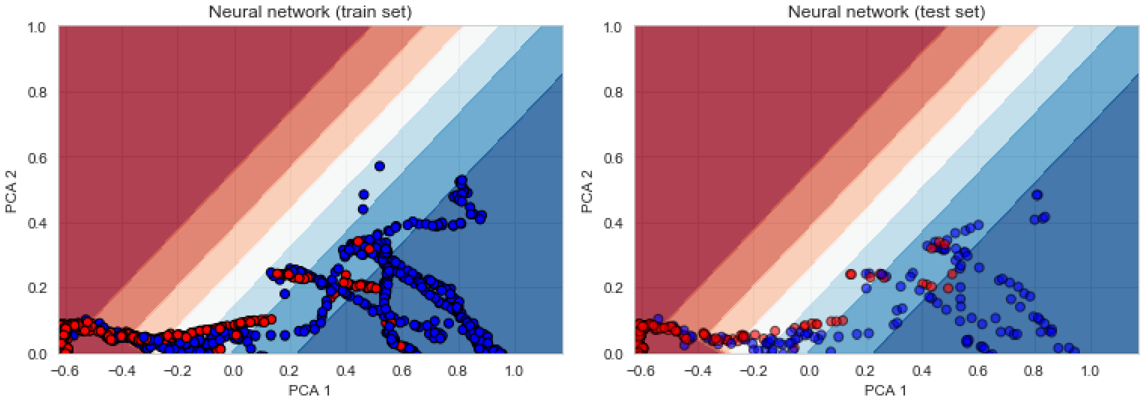


Figure 6: Prediction classification of room occupancy based on Neural Networks (Blue is not occupied, Red is occupied). The accuracy score for this fitness model is: 0.938

For Neural Networks, the grid search determined the most optimal solution with {‘Hidden Layer’: 4,4}, with a confusion matrix of [[332,17] [16,168]]. This matrix shows that the classifier has a few misclassifications on both situation if someone was in the room or not based on the test cases. Of the three machine learning techniques, this method gave the lowest accuracy score which might be due to the low complexity of the fitness function. Further model development would need to be done to get a better predictions using neural networks.

Future Work

In future, we would like to study the following:

1) Apply other factors that might influence occupancy, e.g., current time of day
2) Dimension reduction using Linear Discriminant Analysis for cleaner data and better visualization
3) Improve the Neural Network framework to get a better fitness function for this dataset either by increasing the number of hidden layers, or varying the solver function
4) Use other Classifiers. such as XG boosting, Naïve Bayes and K-nearest neighbour
5) Use *f-test* to predict the model accuracy